Entrega final del proyecto - Aprendizaje no supervisado – MIAD

Gr5: Diana Bayona, Santiago Gutiérrez, Jorge Rodríguez, Juan Pérez

Análisis de asignación de cupos en los colegios distritales mediante técnicas de aprendizaje no supervisado.

Resumen: La gestión de la matrícula en Bogotá es compleja, pues hay una oferta limitada de cupos en cada colegio distrital. El niño, niña, joven o acudiente interesado en obtener un cupo debe diligenciar un formulario de solicitud en el grado y colegio deseado. Cada solicitud recibe un puntaje de acuerdo con las características socioeconómicas del aspirante. Este puntaje determina una priorización de los aspirantes, favoreciendo la asignación del cupo en el colegio solicitado. Teniendo en cuenta lo anterior, el presente trabajo pretende agrupar a los aspirantes en clústeres similares, según sus características socioeconómicas, y comparar los rankings de colegios asignados con diversas métricas, para analizar cómo se distribuyen los cupos y detectar posibles desigualdades. Se propone el uso de algoritmos no supervisados como DBSCAN, K-Means y K-Medoids, que permitan mejorar la equidad al acceso del sistema educativo. Se exploran estrategias para adaptar estos algoritmos, buscando hacerlos más justos y mitigando sesgos al considerar variables sensibles, con el objetivo de lograr resultados de agrupamiento más equitativos.

Introducción

Bogotá está conformada por 20 localidades y 112 Unidades de Planeación Zonal (UPZ), cada una con dinámicas demográficas, sociales y económicas propias, que ocasionan que algunas UPZ concentren mayor demanda del servicio educativo, pero la infraestructura oficial es insuficiente para atenderla en su totalidad, y algunos colegios privados tienen instalaciones deficientes, por lo que el sector público debe impulsar estrategias para que se garantice el acceso a una educación de calidad.

La demanda potencial del sector educativo la determina la población en edad escolar (PEE); con niñas, niños y adolescentes de 5 a 16 años a quienes constitucionalmente les asiste el derecho a acceder a la educación, que solicitan cupos en el sector educativo oficial y no oficial. Sin embargo, existe una población que se encuentra por fuera de este rango de edad: son niños y niñas entre los 3 y 4 años para quienes también se ofrece el servicio educativo, así como a la población en extra - edad, o de adultos, que igualmente demandan cupos en el sistema educativo.

La distribución de la población en edad escolar por localidades para el año 2024 tiene una alta concentración en Suba, Kennedy, Bosa, Ciudad Bolívar, y Engativá, las cuales agrupan el 59,1% de la PEE de la ciudad. Uno de cada tres habitantes de la ciudad está vinculado al sistema educativo, que reúne casi 2,5 millones de personas de los cuales 1,4 millones son niños, niñas y adolescentes en los niveles preescolar, básica, secundaria y media (académica y técnica). El sector oficial atiende a unos 743.000 y el privado a 425.000. Las localidades con mayor matrícula son Suba, Kennedy y Bosa. La oferta del sector oficial está concentrada en las localidades de Bosa, Kennedy y Ciudad Bolívar, donde se atiende cerca del 25% de estudiantes de la ciudad.

La oferta educativa de los colegios del sector oficial se basa en la capacidad de cupos en cada sede educativa, que involucra criterios como la cantidad de aulas disponibles y el tamaño de estas, el número y tamaño de los grupos acorde con el grado, el nivel educativo y la jornada. Al realizar un análisis de oferta y demanda para el sector oficial (en los niveles de preescolar a media), se encuentra que persiste un déficit de cupos en algunas zonas de la ciudad; especialmente UPZ de las localidades de Suba, Kennedy, Bosa, Antonio Nariño, Ciudad Bolívar y Usme. Este déficit de cupos es significativo para el nivel de preescolar en las UPZ El Rincón, Tibabuyes y Patio Bonito.

La asignación de cupos en todos los colegios inicia con la inscripción a través de los canales dispuestos, como la plataforma en línea, los propios colegios y las Direcciones Locales de Educación. Los cupos se asignan mediante un algoritmo que toma en cuenta criterios como la proximidad al lugar de residencia, la disponibilidad en las instituciones educativas y las características poblacionales de los estudiantes (si es víctima del conflicto armado, con discapacidad, pertenece a un grupo étnico). Al final todos estos datos se

traducen en un puntaje que puede influir en la asignación de los cupos. Aunque el objetivo del sistema es promover la equidad, es imperante validar que tan bien esta asignando los cupos, en especial para las poblaciones más vulnerables.

En este contexto, surge la pregunta: **¿Cómo se agrupan los estudiantes de Bogotá según su nivel de vulnerabilidad y qué relación tienen estos grupos con la asignación a colegios mejor ranqueados?** Este interrogante es clave para entender si el sistema realmente está logrando su objetivo de promover la equidad en la educación.

Este es un tema de gran importancia que genera bastante debate, no solo porque toca temas de políticas educativas, sino de cómo combatir la desigualdad y la discriminación. Varios actores tocan estos temas, partiendo desde diferentes perspectivas. La perspectiva de los últimos años es el uso de modelos de aprendizaje no supervisado para la asignación equilibrada, la identificación de patrones ocultos o segmentaciones que no son evidentes a través de los análisis tradicionales.

Amijoyo y Siti Nurhaliza (2023) utilizaron K-means y K-means restringido para zonificar áreas escolares según la distancia entre las escuelas y los domicilios de los estudiantes, concluyendo que K-means restringido produce clústeres más equilibrados, lo que facilita una distribución justa de recursos. Cahapin et al. (2023) emplearon K-means, Hierarchical y DBSCAN para agrupar estudiantes en función de 11 características, proponiendo que estos algoritmos pueden mejorar las decisiones educativas. Le Quy et al. (2023) aplicaron diversas técnicas de clustering, como K-means y Fuzzy C-means, para agrupar estudiantes de manera equitativa, subrayando la importancia de evitar sesgos. Lázaro y Núñez (2023) desarrollaron modelos para ofrecer tutorías focalizadas en la Universidad Nacional de Ingeniería, destacando la Propagación de Afinidad como la técnica más eficaz. Finalmente, Le, Friege y Ntoutsi (2023) revisaron la literatura sobre clustering en educación, explorando cómo estos métodos pueden adaptarse para mitigar sesgos y promover la equidad.

Materiales y Métodos

Para realizar el ejercicio de agrupación de este trabajo, se utilizó como principal insumo la base de datos de solicitudes de cupos nuevos en colegios distritales de la Secretaría de Educación del Distrito para la vigencia 2024, que contiene 86.815 solicitudes y 137 características. Como características se tiene información como edad, dirección, localidad, barrio, genero, etnia, discapacidad, si es víctima de conflicto, talento de los aspirantes, entre otras. La información cuenta con el estado de la solicitud y el grado y colegio deseado. Se identifica que, de las 137 características, varios campos se encuentran vacíos o no cumplen con un porcentaje adecuado de completitud, tienen información redundante, son datos de identificación del aspirante o datos de identificación de los familiares. Al final, información que no aporta al análisis de este trabajo. Por lo cual, se depura la base y quedan 27 características. Para complementar el trabajo, se adicionan otras bases de datos como la capacidad instalada y el puntaje de ICFES de los colegios, así como los cupos ofrecidos por grado de cada colegio.

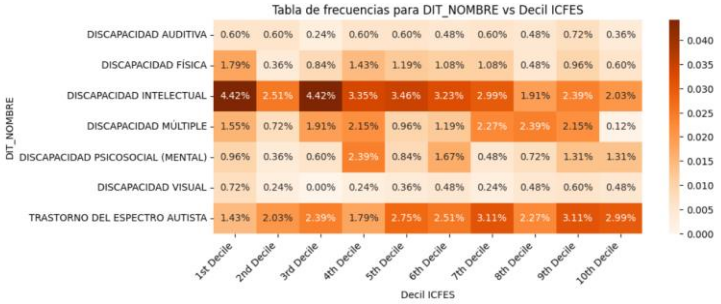
Dentro del análisis descriptivo, encontramos que la mayoría de los aspirantes buscan cupo en preescolar (33.443) y básica primaria (30.139). Las localidades que presentan mayor concentración de los aspirantes son Suba (12.229), Bosa (13.458), Kennedy (12.997) y Ciudad Bolívar (12.031), capturando aproximadamente el 58,42 % del total de aspirantes. Del total de aspirantes, solo el 14,96%, equivalente a 12.991 aspirantes, cuentan con algún talento excepcional. El talento más destacado dentro de estos aspirantes es el talento en actividades física, ejercicio y deporte con 7.525 aspirantes. Por otro lado, los talentos en artes o letras, liderazgo social y emprendimiento, y tecnología cuentan con 3.547, 1.071 y 943 aspirantes respectivamente.

En las variables que describen si algún estudiante presenta condiciones de vulnerabilidad como pertenecer a alguna etnia, ser víctima de conflicto armado o presentar alguna discapacidad, la mayoría no las presenta. Sin embargo, entre las discapacidades que más presentan los aspirantes se encuentra discapacidad intelectual (392) y trastorno del espectro autista (355). Así mismo, 4.598 estudiantes fueron víctimas del conflicto armado. Mientras que la etnia que más se presenta es la de afrodescendiente con 1.239 jóvenes, seguida de la categoría de otros (247), wayuu (201) y pijao (120). Aunque la mayoría de los estudiantes no

presenta condiciones de vulnerabilidad, existe una notable presencia de ciertos grupos vulnerables que se ven expuestos a enfrentar desafíos en su contexto social.

Los talentos excepcionales en diferentes áreas a lo largo de los deciles de desempeño en el examen ICFES. Los talentos en Artes, Liderazgo Social y Tecnología tienden a concentrarse en los deciles superiores, alcanzando sus picos en el 9° y 10° decil. En cambio, los talentos en Ciencias Naturales y Sociales presentan porcentajes bajos en todos los deciles, aunque también aumentan ligeramente en los niveles superiores. El talento en Actividad Física muestra una distribución más uniforme, con un aumento notable en los deciles altos. Sorprendentemente, esto va en contra de la intuición, ya que se esperaría que los estudiantes con talentos excepcionales estuvieran más concentrados en los deciles superiores de rendimiento.

Ilustración 1 – Distribución de Discapacidad por Decil del Puntaje ICFES del Colegio Distrital



Fuente: Creación propia

La tabla muestra la distribución de estudiantes con distintas discapacidades a lo largo de los deciles de desempeño en el examen ICFES, revelando variaciones significativas según el tipo de discapacidad. Mientras que estudiantes con discapacidad auditiva y visual están distribuidos de manera más uniforme, aquellos con discapacidades intelectuales o físicas se concentran en los deciles inferiores, lo que sugiere mayores dificultades para alcanzar altos niveles de rendimiento. Esto plantea importantes desafíos para una asignación justa de cupos en colegios, destacando la necesidad de ajustar los criterios de selección para mitigar posibles desventajas y garantizar un acceso equitativo para todos los estudiantes, independientemente de sus limitaciones.

En este proyecto, proponemos una metodología que convierte variables categóricas en texto, el cual se transforma en vectores numéricos mediante embeddings utilizando un modelo de lenguaje (LLM). A continuación, aplicamos técnicas de reducción dimensional, como PCA, y ejecutamos el algoritmo de clustering con datos numéricos. Se sabe que los embeddings ayudan a que frases y textos con significados semánticos similares se representen como vectores cercanos, lo cual puede beneficiar la implementación de algoritmos como el DBSCAN donde la densidad de los datos juega un papel importante.

Se planea implementar DBSCAN como algoritmo principal, dado que es especialmente adecuado para manejar bases de datos desbalanceadas y con valores atípicos. DBSCAN se basa en la densidad de los puntos y permite identificar grupos de estudiantes con características similares sin requerir que se especifique el número de clúster previamente, además de ser más robusto frente a outliers, lo cual es crucial en un conjunto de datos desbalanceado. Los embeddings generados a partir de un modelo de lenguaje (LLM) también benefician a DBSCAN, ya que representan textos con significados semánticos similares como vectores cercanos, mejorando la agrupación basada en densidad. Como alternativas, se usarán los algoritmos de K-Means y K-Medoids. En resumen, se implementan los siguientes algoritmos: **LLM Total + PCA + DBSCAN, LLM Total + PCA + K-Means y LLM Total + PCA + K-Medoids.**

Resultados y Discusión

A través de diversas técnicas analíticas, incluyendo el Análisis de Componentes Principales (PCA) y los diferentes algoritmos de clustering, se han identificado patrones y relaciones entre los datos. A continuación, se presentan los hallazgos clave que emergen de este análisis, acompañados de interpretaciones que destacan su relevancia en el contexto de la investigación.

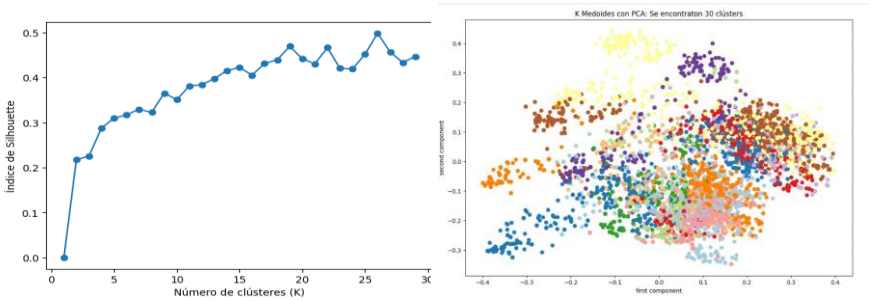
Inicialmente se desarrolló el modelo de DBSCAN, el cual es excelente para detectar clústeres de forma arbitraria y puede identificar outliers como ruido. Sin embargo, puede tener problemas con la configuración de los parámetros eps y min_samples. Para este caso, el parámetro de eps no tuvo una adecuada estimación dado que al estar los datos de embeddings normalizados las distancias eran demasiado pequeñas, lo que hizo que las variaciones fueran mínimas. Por ende, se recurrió a seleccionar el eps de forma manual. Siendo así, se seleccionó un eps de 1.2 fijo, y a partir de este se determinó la cantidad de min_samples, de esta manera para un eps de 1.2 y min_samples de 39 se generaron finalmente 63 clusteres.

K-Means agrupa los datos en un número predefinido de clústeres (k). El algoritmo divide los puntos en grupos de manera que cada grupo tiene un “centro” y los puntos más cercanos a ese centro pertenecen al mismo clúster. Es rápido y fácil de usar, pero no maneja bien los valores atípicos y los clústeres deben tener una forma más o menos redonda. Pide definir el número de clústeres, para seleccionar el numero optimo se utiliza o la varianza intra cluster o el coeficiente de siluette.

K-Medoids usa puntos reales del conjunto de datos como centros de clúster. Esto lo hace más robusto ante valores atípicos. El número de clúster se determina utilizando el criterio de Silhouette. Al analizar las características de los clústeres generados por K-Medoides, se identificaron patrones coherentes que se alinean con los resultados obtenidos mediante DBSCAN y K-Means, lo que facilita una interpretación más clara y menos propensa al sobreajuste de los datos.

Dentro de los algoritmos que se implementaron, el que menos clústeres genero fue el algoritmo de K-Medoides con 26 grupos. Mientras que los algoritmos de K-means y DBSCAN generaron 79 y 63 clústeres respectivamente. Debido a esto se considera la implementación de LLM Total + PCA + K-Medoids como la mejor dado que con un menor número de clústeres, la interpretación de los resultados se vuelve más sencilla. Es más fácil entender y comunicar los hallazgos cuando hay menos grupos que analizar.

Aunque el mejor modelo es el K Medoids, este puede presentar algunas limitaciones como su baja escalabilidad, siendo menos eficiente que K-Means en grandes conjuntos de datos debido al cálculo de distancias entre todos los puntos. Requiere especificar el número de clústers previamente, lo que puede ser complicado si no se conoce el valor óptimo. Su rendimiento disminuye con datos de alta dimensionalidad y su convergencia puede ser lenta, quedando atrapado en óptimos locales. Aunque es más robusto ante outliers que K-Means, no es inmune a ellos. Estas limitaciones lo hacen menos adecuado para grandes volúmenes de datos o situaciones complejas sin ajustes adicionales.



A partir de los resultados obtenidos con el algoritmo de k-medoides se llegan a las siguientes observaciones:

- Los aspirantes presentan una diversidad educativa que abarca desde preescolar hasta educación media, con mayor concentración en primaria y secundaria. Aquellos con talentos excepcionales son asignados a colegios con deciles más altos, obteniendo mejores condiciones educativas y, aparentemente, mejores resultados académicos.
- La distribución étnica es equitativa a lo largo de los deciles del ICFES, lo que indica que la etnia no es un factor discriminatorio en la asignación de cupos. Este patrón persiste incluso al analizar

niveles secundarios dentro de cada clúster, lo que refuerza la idea de que la pertenencia étnica no influye en el proceso.
- Finalmente, el uso de 26 clústeres plantea el riesgo de sobreajuste o captura de ruido, lo que sugiere que la segmentación puede estar reflejando variaciones irrelevantes, en lugar de patrones significativos.

Dentro de las recomendaciones para futuros análisis se sugiere mejorar la selección de variables y el análisis de datos desbalanceados mediante técnicas como XGBoost y modelos avanzados como HDBSCAN y OPTICS. También proponen aplicar regularización y utilizar métodos de clustering por consenso para obtener agrupaciones más representativas.

Conclusión

El análisis revela que los aspirantes con talentos excepcionales son asignados a colegios con mejor desempeño, lo que les permite acceder a condiciones educativas más favorables y obtener mejores resultados en pruebas académicas. Además, la distribución de etnias es equitativa a lo largo de los diferentes deciles de desempeño del ICFES, lo que indica que la etnia no influye como factor de discriminación o vulnerabilidad en el proceso de asignación escolar.

Sin embargo, el algoritmo arrojo un número elevado de clústeres, lo que sugiere que podría estar sobre ajustándose a los datos, es decir, agrupando los puntos de manera excesivamente específica en lugar de identificar patrones generales más representativos. Esto puede llevar a una pérdida de capacidad de generalización e interpretación, dificultando la extracción de conclusiones mas útiles y prácticas.

Este análisis preliminar nos da una primera aproximación a la estructura subyacente de los datos, pero no es suficiente para realizar interpretaciones definitivas o totalmente confiables. Será necesario ajustar los parámetros de los algoritmos y explorar enfoques adicionales para mejorar la calidad y relevancia de los clústeres, así como realizar una validación más robusta de los resultados.

Bibliografía

- Amijoyo, T., & Siti Nurhaliza, M. (2023). Application of K-means clustering algorithm method in new student admissions. *Jurnal Inovatif: Inovasi Teknologi Informasi dan Informatika*, *6*(1), 92–100. https://ejournal.uika-bogor.ac.id/index.php/INOVA-TIF/article/view/14788
- Cahapin, E., Malabag, B., Santiago Jr, C., Reyes, J., Legaspi, G., & Adrales, K. (2023). Clustering of students admission data using K-means, hierarchical, and DBSCAN algorithms. *Bulletin of Electrical Engineering and Informatics*, *12*(6), 3647–3656. https://doi.org/10.11591/eei.v12i6.4849
- Lazaro, E., & Nuñez, Y. (2023). Segmentation of university students using clustering and considering a virtual cycle. En *Proceedings of the Latin American and Caribbean Conference for Engineering and Technology (LACCEI)*. https://laccei.org/LACCEI2023-BuenosAires/papers/Contribution_1355_a.pdf
- Le Quy, T., Friege, G., & Ntoutsi, E. (2023). A review of clustering models in educational data science towards fairness-aware learning. *arXiv*. https://arxiv.org/abs/2301.03421
- Le, T., Friege, G., & Ntoutsi, E. (2023). A review of clustering models in educational data science towards fairness-aware learning. *L3S Research Center, Leibniz University Hannover*. https://arxiv.org/pdf/2301.03421
- Secretaría Distrital de Educación. (2024). Bases de diagnóstico para Plan Distrital de Desarrollo 2024- 2028.

GitHub: El repositorio del proyecto actual se encuentra en el siguiente link.