Clusterización del Operativo de Evaluación de la Calidad Educativa “Aprender 2018” del Sexto Grado de las Escuelas Primarias del país

Andrés Francisco Farías[2], Germán Antonio Montejano[3],

Ana Gabriela Garis[3], Sebastián Javier Farías[2], Andrés Alejandro Farías[2]

[2]Universidad Nacional de La Rioja, La Rioja, Argentina
afarias665@yahoo.com.ar, sebajavfarias@gmail.com
andres_af86@hotmail.com
[3]Universidad Nacional de San Luis, San Luis, Argentina
gmonte@unsl.edu.ar, agaris@unsl.edu.ar

Abstract. En este trabajo se expone la aplicación de técnicas de Clusterización del Operativo de Evaluación de la Calidad Educativa “Aprender 2018” del Sexto Grado de las Escuelas Primarias del país. Donde el análisis se realizó sobre dos variables: Puntaje en Lengua y Puntaje en Matemáticas, tomando todos los alumnos que participaron del Operativo. Para luego procesar los datos obtenidos de cada Cluster utilizando modelos de Machine Learning, que permitiera comprender sus características. El procedimiento siguió los siguientes pasos: determinación del número óptimo de cluster, determinación gráfica de la distribución de los Clusters, uso de Machine Learning e interpretación de los resultados.

1 Introducción

El punto de partida es el dataset del Operativo de Evaluación de la Calidad Educativa “Aprender 2018” [1, 2] del Sexto Grado de las Escuelas Primarias del país, de donde se obtuvieron los datos sobre los que se realizará el proceso de clusterización.

Se seleccionan dos variables para analizar su relación y como parámetros para determinar los clusters, estas fueron:

- Puntaje en Lengua
- Puntaje en Matemáticas

En la selección se adoptaron los valores que correspondían a todos los alumnos que participaron.

Como procedimiento metodológico para este trabajo, para completar los valores faltantes, se utilizó la mediana de cada columna.

2 Antecedentes

Hay algunas experiencias que pueden servir como antecedentes para el uso de las herramientas de Clusterización, pero concretamente en el ámbito educativo, y más específicamente aún, analizando las avaluaciones Aprender, encontramos la investigación realizada por Pizarro Levi & Antequera (2026) denominada Desigualdades educativas intra-provinciales: un análisis de cluster para la provincia de La Rioja, Argentina (2022). [3, 4]

El estudio de Pizarro Levi y Antequera (2026) constituye uno de los aportes más recientes y exhaustivos sobre el uso de análisis de cluster jerárquicos para caracterizar desigualdades educativas. Los autores analizan los 18 departamentos de la provincia de La Rioja mediante un enfoque multidimensional, integrando datos de APRENDER 2022, el Censo 2022 y registros administrativos provinciales.

El estudio busca identificar patrones territoriales de desigualdad educativa en cuatro dimensiones: Acceso y permanencia, Calidad y rendimiento, Entorno educativo y Condiciones socioeconómicas.

Para ello, aplican: Cluster jerárquicos con método de Ward y distancia euclidiana, Correlación cofenética e índice de Davies-Bouldin para evaluar la calidad del agrupamiento y Prueba de Kruskal-Wallis para detectar heterogeneidad intra-cluster.

Este trabajo es especialmente relevante como antecedente ya que demuestra la pertinencia del análisis de cluster para estudiar, entre otras variables, las desigualdades educativas entendida como factores asociados o condicionantes de los aprendizajes, y ofrece un marco metodológico replicable para investigaciones posteriores.

Otro de los antecedentes, en este caso a nivel nacional es el de Farías, Durán y Figueroa, 2008 denominado *Las Técnicas de Clustering en la Personalización de Sistemas de e-Learning.*[1]

Esta investigación propone ofrecer una solución al abandono de estudiantes de carreras a distancia o e-learning, por lo que identifica como problemas relacionados con la no adaptación de las aulas virtuales o plataformas educativas a los diferentes estilos de aprendizaje de los usuarios.

En esta investigación se concluye que los sistemas de e-learning actuales presentan importantes limitaciones para adaptarse a las características individuales de los estudiantes, lo que repercute negativamente en la asimilación de los contenidos y en la calidad del aprendizaje.

Frente a esta problemática, el artículo propone el uso de técnicas de análisis de cluster como una alternativa para identificar el estilo de aprendizaje dominante de cada alumno y así posibilitar la personalización de los contenidos y de las estrategias de enseñanza según sus necesidades y preferencias.

---

[1] Farías, R. A., Durán, E. B., & Figueroa, S. G. (2008). Las técnicas de clustering en la personalización de sistemas de e-learning. En Actas del CACIC 2008. Universidad Nacional de Santiago del Estero. https://sedici.unlp.edu.ar/handle/10915/21990

3 Clusters

El clustering o análisis de conglomerados constituye una de las técnicas fundamentales del análisis de datos y del aprendizaje no supervisado, cuyo objetivo principal es la identificación de estructuras latentes y patrones intrínsecos dentro de conjuntos de datos sin etiquetas previas. En el ámbito de la investigación de datos, la técnica en cuestión organiza los datos en grupos o clusters. Estos grupos se caracterizan por la presencia de una mayor similitud entre los elementos pertenecientes a un mismo grupo que a los pertenecientes a otros grupos. Esta organización se lleva a cabo mediante una métrica de distancia o similitud previamente definida. La capacidad de descubrir conocimiento inherente al clustering lo posiciona como una herramienta fundamental para el análisis exploratorio de datos en contextos donde no se cuenta con información previa sobre la estructura de los mismos.[5, 6]

4 Métodos de agrupación en Aprendizaje Automático

Los métodos de agrupación en Aprendizaje Automático son de vital importancia para la organización y clasificación de datos en grupos o clústeres basados en similitudes o patrones [7, 8]. A continuación, se presentan algunos de los métodos más empleados:

- **Algoritmo de Mean Shift**: un procedimiento matemático de vanguardia, identifica modas en conjuntos de datos mediante la búsqueda de regiones densas.
- **Clustering basado en centroides:** es un método de agrupación de datos que se fundamenta en la distancia entre los centroides de los datos. Este método divide un conjunto de datos en grupos similares según la proximidad entre sus centroides.
- **DBSCAN**: es un algoritmo que determina automáticamente el número de clusters y es efectivo en la detección de outliers.
- **K-Means:** es un método que se utiliza para la inicialización de un número de clusters y la agrupación de los datos en grupos que contengan a los puntos de datos más cercanos.

5 K-means: metodología para la clusterización de datos

5.1 Selección de variables y gráfico de relación

Como se explicó en la introducción, con las variables adoptadas se realizó un gráfico de puntos, en Figura 1:

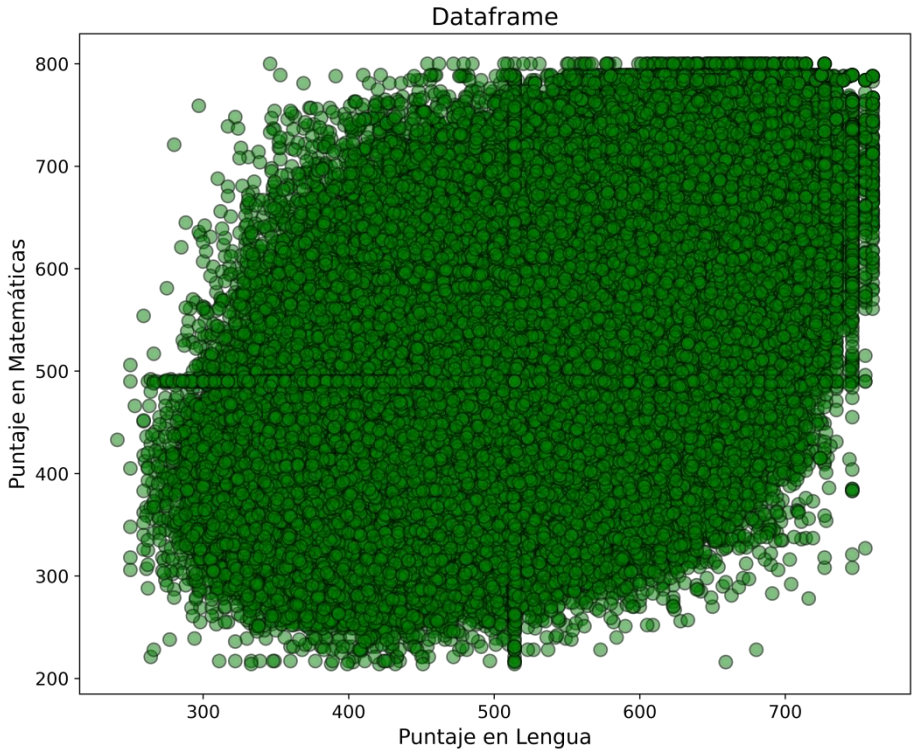

Fig. 1. Relación de variables

5.2 Determinación de número de clusters

Utilizamos dos técnicas gráficas para determinar el número de clusters [9, 10]:

- Método de Elbow
- Método de la Silueta

En los dos procedimientos el número óptimo de varía de dos a cuatro clusters, a continuación, en Figura 2 tenemos Método de Elbow con números óptimos que varian de dos a cuatro custers y en Figura 3 el Método de la Silueta cuyo número óptimo es de dos clusters.

:

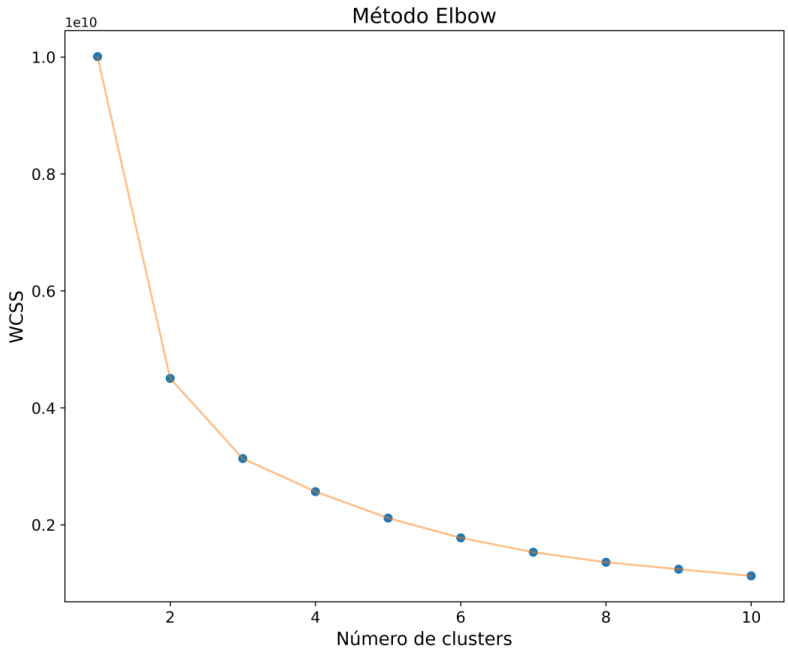

Fig. 2. Método de Elbow

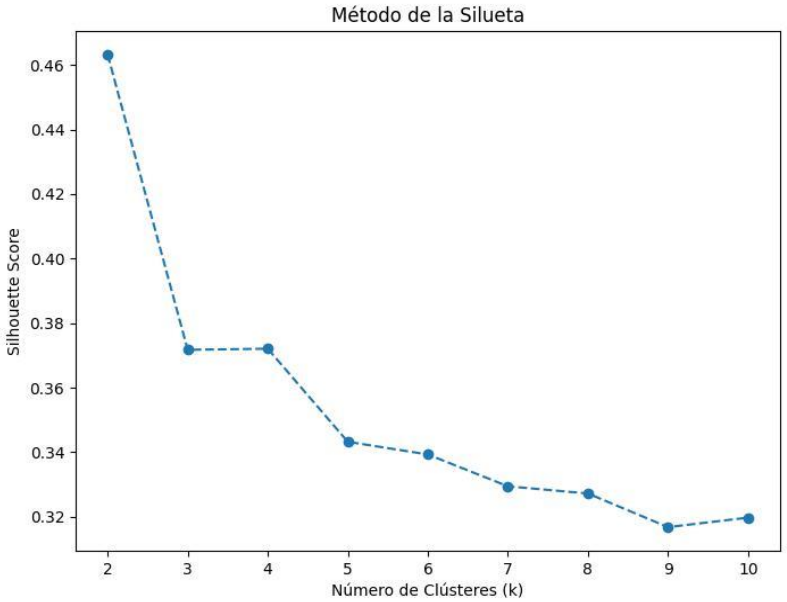

Fig. 3. Método de la Silueta

5.3 Adopción de una clusterización para analizar

Los resultados observados en las figuras anteriores nos indican la cantidad más apropiada de clusters para los datos en estudio, que varían de dos a cuatro agrupamientos.

Para este estudio, se adopta para el análisis la distribución de cuatro clusters.

5.4 Cuatro clusters

Realizados los cálculos se visualiza los resultados la clusterización, en Figura 4:

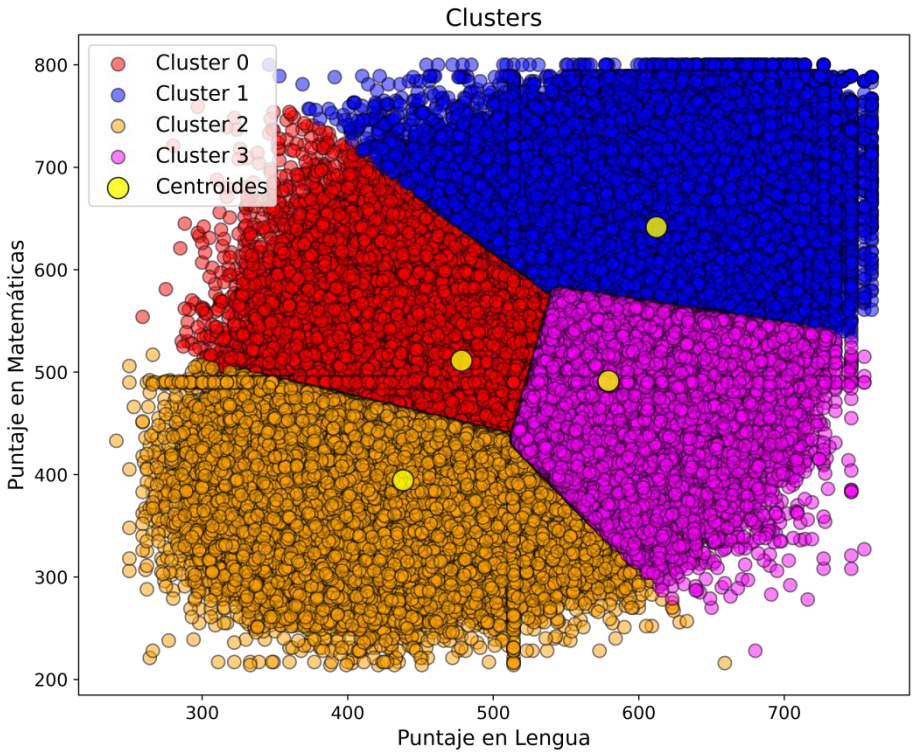

Fig. 4. Distribución de cuatro clusters

5.5 Selección de variables

En Tabla 1, se indican las variables utilizadas en el análisis

Tabla 1. Selección de variables

Variable	Descripción
¿Personas/habitación?	Personas por habitación, indica el nivel de hacinamiento

Sector	Sector de gestión: pública o privada
Trabajas?	Trabajo dentro o fuera del hogar
Libros	Cantidad de libros
¿Nivel educativo de tus Padres?	Nivel educativo de la madre y el padre
Sexo	Sexo
Discriminación	Todos los niveles de discriminación
¿Jardín de infantes?	Concurrencia al Jardín de infantes
Edad	Edad
¿Repetiste de grado?	Repitencia
Ámbito	Ámbito rural o urbano

5.6 Importancia de las variables en el Puntaje en Lengua por cluster

A partir de los puntajes en Lengua se analizan la importancia de las otras variables en cada uno de los clusters.

Cluster 0: Puntajes Medios/Altos en Matemática y Medios/Bajos en Lengua

En la figura 5:

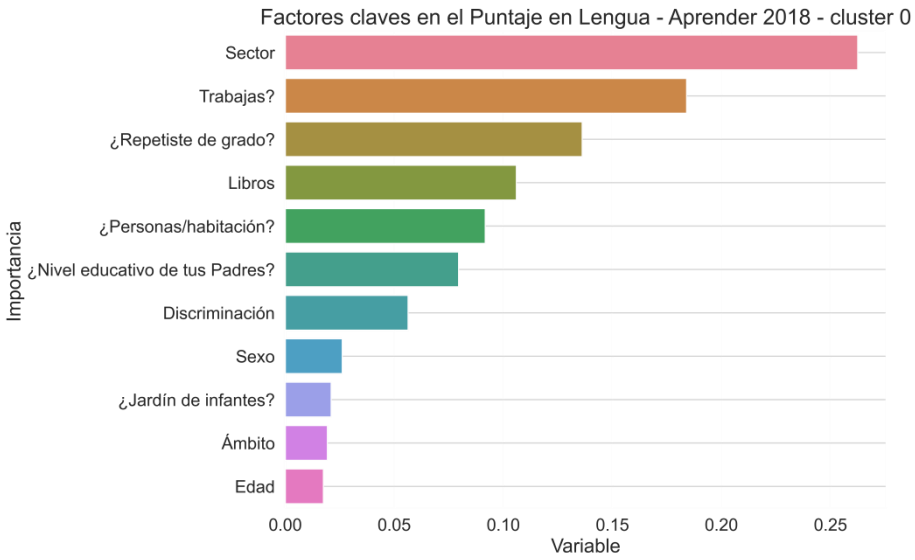

Fig. 5. Importancia de las variables, cluster 0

Cluster 1: Puntajes Altos en Matemática y en Lengua

En la figura 6:

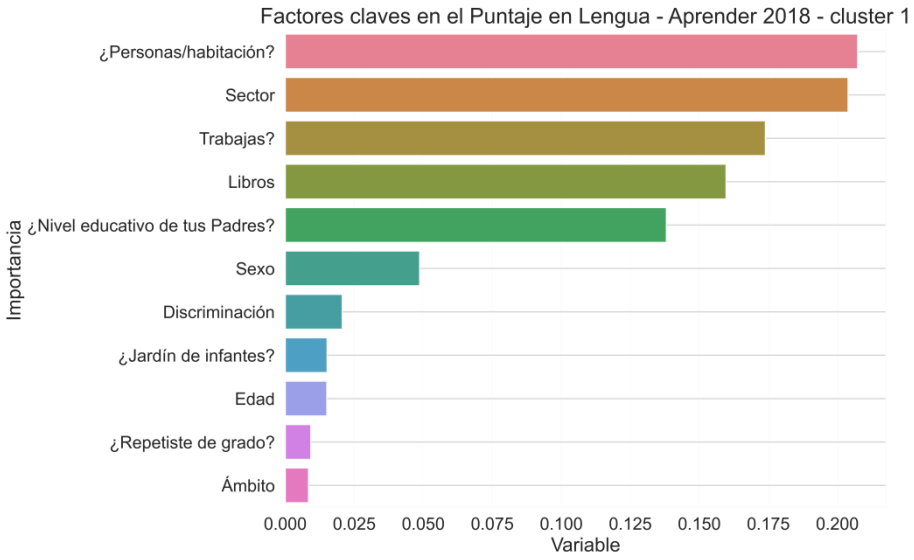

Fig. 6. Importancia de las variables, cluster 1

Cluster 2: Puntajes Bajos en Matemática y en Lengua

En la figura 7:

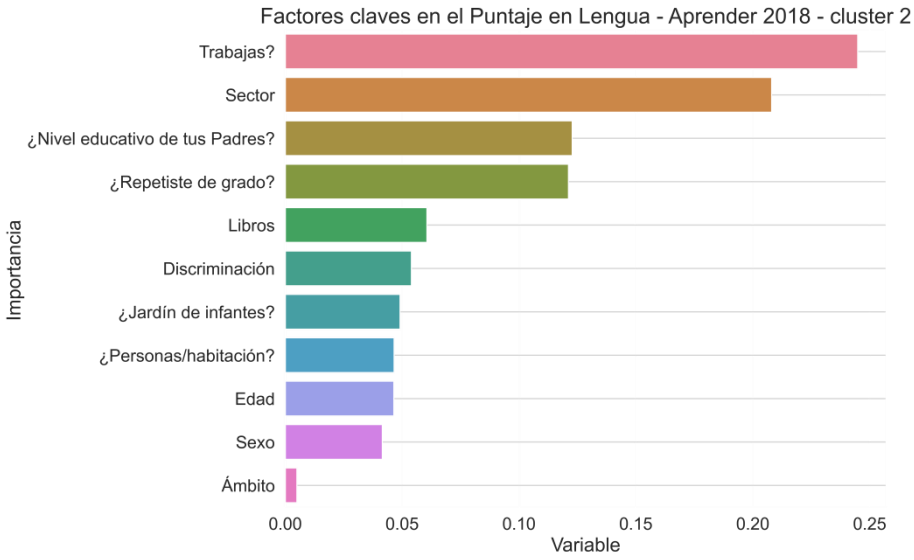

Fig. 7. Importancia de las variables, cluster 2

Cluster 3: Puntajes Medios/Bajos en Matemática y Medios/Altos en Lengua

En la figura 8:

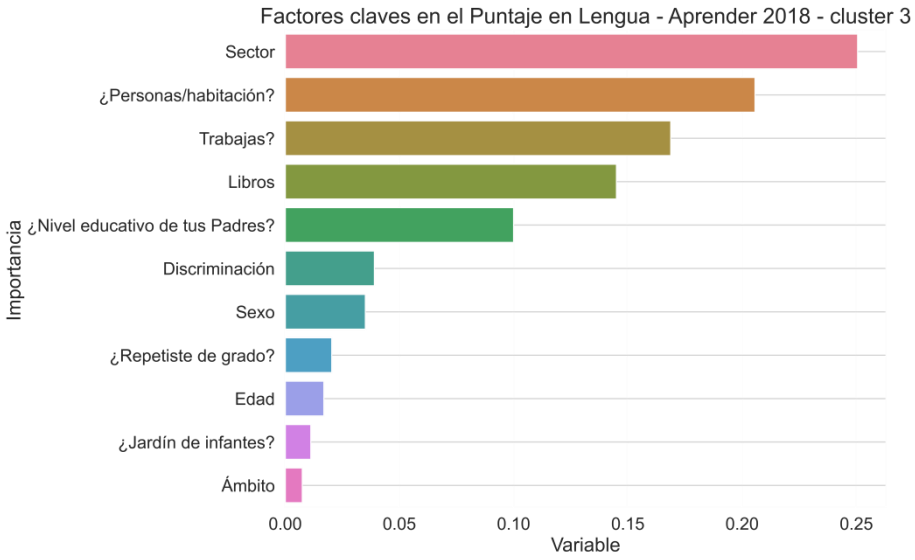

Fig. 8. Importancia de las variables, cluster 3

5.7 Análisis de los resultados

En Tabla 2:

Tabla 2. Análisis de los resultados

Variable	Cluster 0	Cluster 1	Cluster 2	Cluster 3
¿Personas/habitación?	0,092	0,207	0,047	0,206
Sector	0,262	0,204	0,208	0,251
Trabajas?	0,184	0,174	0,245	0,169
Libros	0,106	0,160	0,061	0,145
¿Nivel educativo de tus Padres?	0,079	0,138	0,123	0,100
Sexo	0,026	0,049	0,042	0,035
Discriminación	0,056	0,021	0,054	0,039
¿Jardín de infantes?	0,021	0,015	0,049	0,011
Edad	0,018	0,015	0,046	0,017
¿Repetiste de grado?	0,136	0,009	0,121	0,020
Ámbito	0,019	0,008	0,005	0,007

En la figura 9:

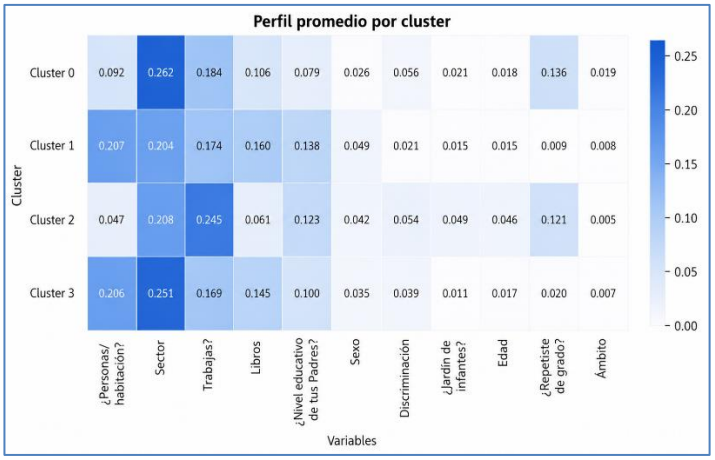

Fig. 9. Comparación de variables por cluster

5.8 Importancia de las variables en el Puntaje en Matemáticas por cluster

A partir de los puntajes en Lengua se analizan la importancia de las otras variables por cada uno de los clusters.

Cluster 0: Puntajes Medios/Altos en Matemática y Medios/Bajos en Lengua

En la figura 10:

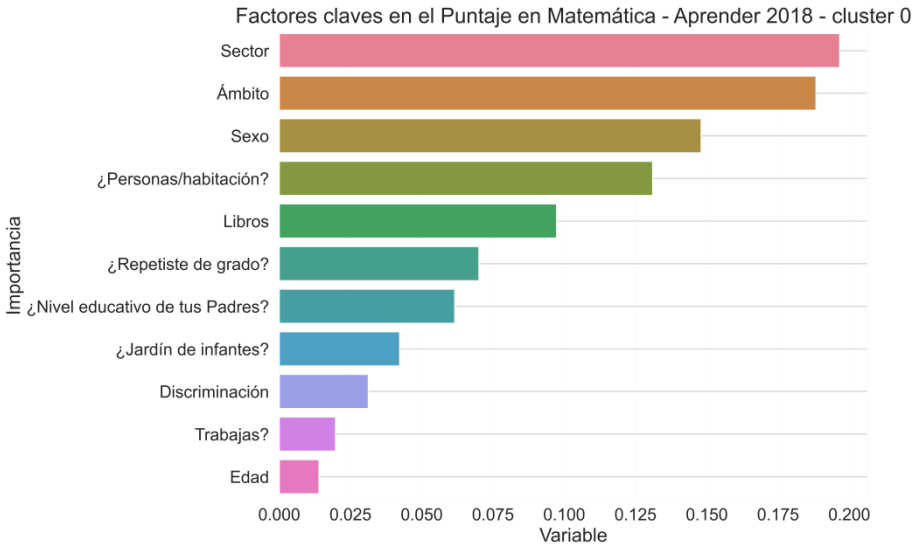

Fig. 10. Importancia de las variables, cluster 0

Cluster 1: Puntajes Altos en Matemática y en Lengua

En la figura 11:

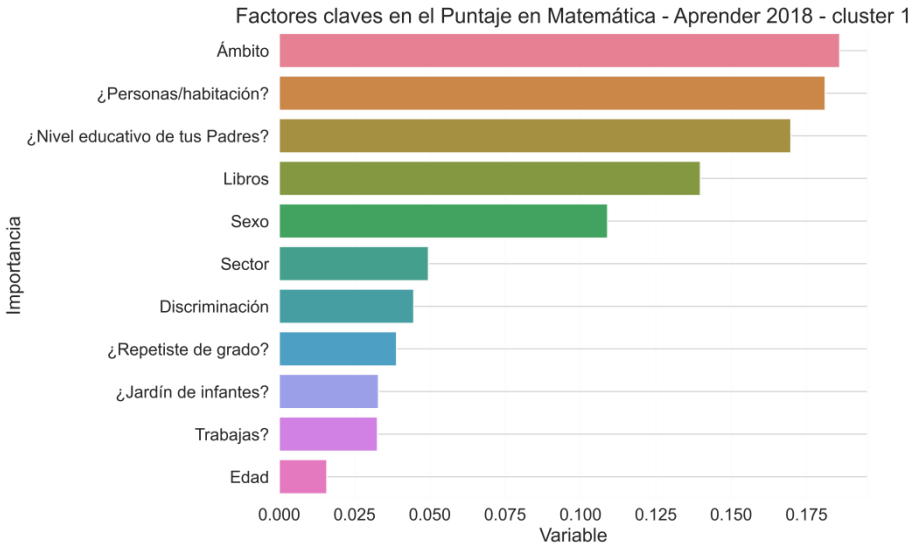

Fig. 11. Importancia de las variables, cluster 1

Cluster 2: Puntajes Bajos en Matemática y en Lengua

En la figura 12:

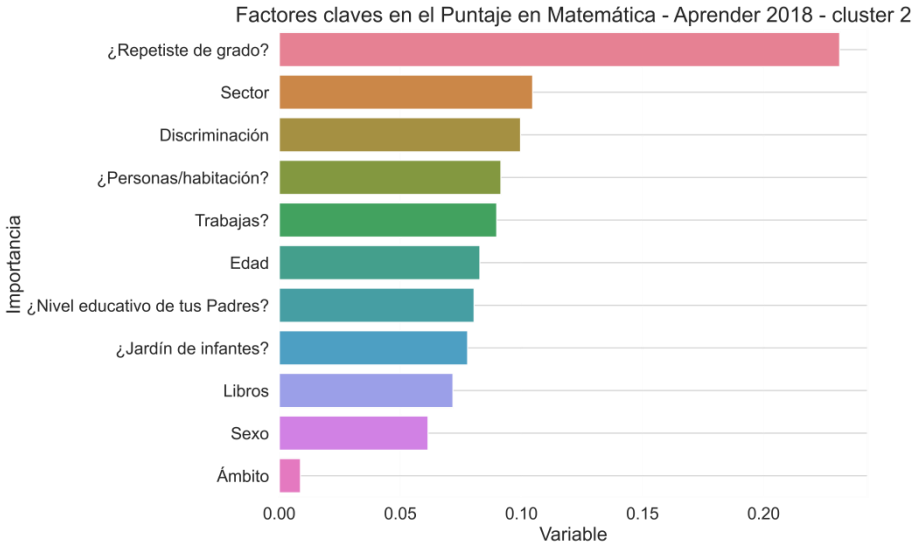

Fig. 12. Importancia de las variables, cluster 2

Cluster 3: Puntajes Medios/Bajos en Matemática y Medios/Altos en Lengua

En la figura 13:

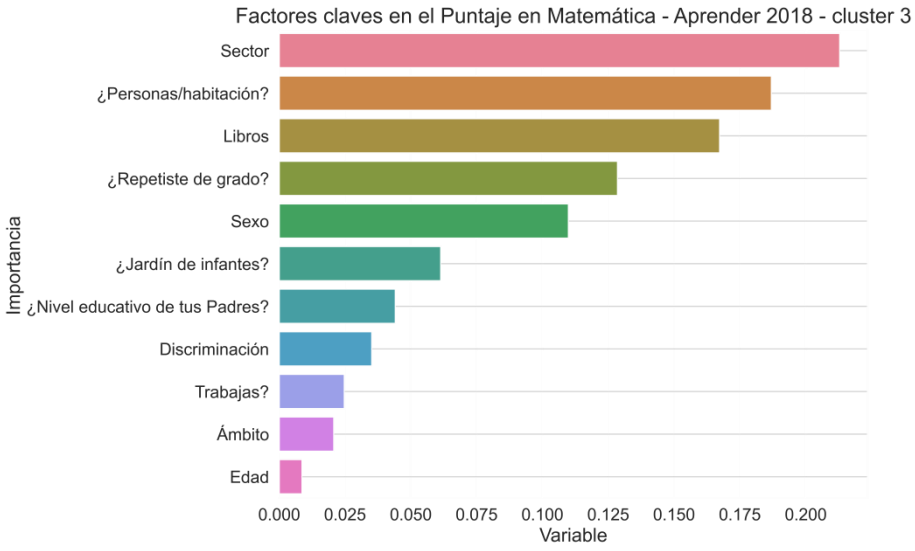

Fig. 13. Importancia de las variables, cluster 3

5.9 Análisis de los resultados

En Tabla 3:

Tabla 3. Análisis de los resultados

Variable	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Ámbito	0,188	0,186	0,009	0,021
¿Personas/habitación?	0,131	0,181	0,092	0,187
¿Nivel educativo de tus Padres?	0,062	0,170	0,080	0,044
Libros	0,097	0,140	0,072	0,167
Sexo	0,148	0,109	0,061	0,110
Sector	0,197	0,050	0,105	0,213
Discriminación	0,031	0,045	0,100	0,035
¿Repetiste de grado?	0,070	0,039	0,231	0,128
¿Jardín de infantes?	0,042	0,033	0,078	0,061
Trabajas?	0,020	0,033	0,090	0,025
Edad	0,014	0,016	0,083	0,009

En Figura 14:

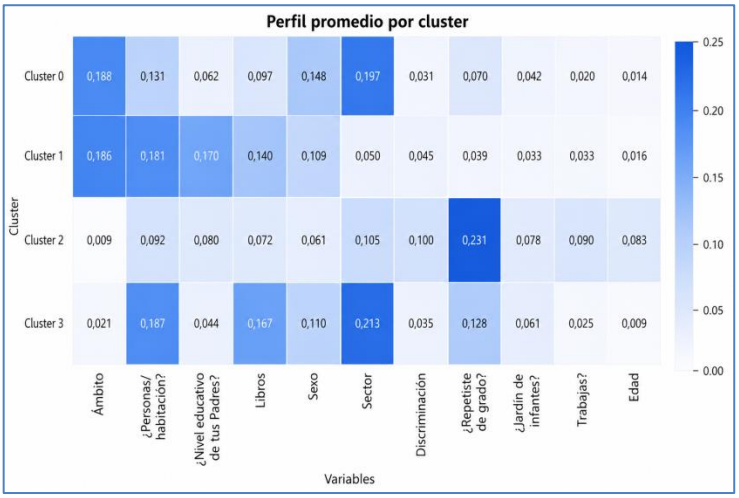

Fig. 14. Comparación de variables por cluster

Con respecto a las variables que mayor asociación tienen con Matemática, se observa que el Ámbito tiene una clara predominancia en los Clusters que referencian mayores puntajes en Matemática, mientras que tienes muy bajo nivel de asociación los clusters de puntajes bajos en esta área. Mientras que la variable Repitencia evidencia mayores asociaciones cuando se realiza el análisis de los Clusters que implican menores desempeños en el área de matemática.

El análisis de la variable de Hacinamiento tiene alta asociación cuando se la trabaja con los desempeños Altos y Bajos, y un poco menos con los desempeños medios, situación similar a la variable Posesión de Libros.

Por último es destacable la variable Sector y el peso superior que tiene en los clusters de desempeños intermedios, Mientras que baja considerablemente cuando los agrupamientos se ubican en los extremos de altos y bajos desempeños en ambos espacios.

6 Conclusiones

La técnica de clusterización del Machine Learning, en este caso el método Kmeans, se pudo implementar sin ninguna dificultad en los resultados del Aprender 2018. los resultados permiten sostener que la **clusterización** resultó adecuada para el propósito exploratorio del estudio, en tanto que posibilitó distinguir **cuatro agrupamientos** de estudiantes a partir de sus puntajes en **Lengua** y **Matemática**, con centroides definidos de manera consistente.

Sobre esta base, el aporte específico del Machine Learning en este trabajo consiste en agregar un **nivel adicional de lectura.** Mientras el Clustering, permite observar

perfiles combinados de desempeño Lengua-Matemática (por ejemplo, alto-alto; alto-bajo; bajo-alto; bajo-bajo) y, a partir de allí, explorar si los **factores asociados** operan de forma diferencial según el tipo de perfil, el Machine Learning, vuelve a poner el foco en los factores asociados a cada una de las disciplinas a finde profundizar su interpretación en cada una de ellas.

En esa línea se puede realizar la siguiente lectura de los datos obtenidos.

En el área de Lengua hay variables que claramente tienen mayor y menor peso de asociación con los desempeños, como lo es el caso del Sector de Gestión, donde en todos los Cluster obtiene altos niveles de correlación, principalmente en los Cluster 0 y 3 que son los referidos a puntajes intermedios. mientas que el caso del Ámbito Urbano o Rural para todos los agrupamientos es notorio el bajo nivel de asociación. Puede incluirse también al Sexo, la Discriminación, Edad y Repitencia como variables que en general se asocian muy débilmente con los desempeños en Lengua.

También es interesante notar cómo el hacinamiento tiene mayor peso en los clusters de los extremos buenos y malos desempeños generales, mientras que es notoria la baja asociación cuando los promedios son medios en ambos espacios. Similar situación con la posesión de Libros en el Hogar. Se mantiene con una estable relación en todos los clusters el Nivel Educativo de Madre y Padre en Lengua.

Con respecto a las variables que mayor asociación tienen con Matemática, se observa que el Ámbito tiene una clara predominancia en los Clusters que referencian mayores puntajes en Matemática, mientras que tienes muy bajo nivel de asociación los clusters de puntajes bajos en esta área. Mientras que la variable Repitencia evidencia mayores asociaciones cuando se realiza el análisis de los Clusters que implican menores desempeños en el área de matemática.

El análisis de la variable de Hacinamiento tiene alta asociación cuando se la trabaja con los desempeños Altos y Bajos, y un poco menos con los desempeños medios, situación similar a la variable Posesión de Libros.

Por último, es destacable la variable Sector y el peso superior que tiene en los clusters de desempeños intermedios, Mientras que baja considerablemente cuando los agrupamientos se ubican en los extremos de altos y bajos desempeños en ambos espacios.

En suma, los resultados del presente trabajo respaldan que la clusterización no solo fue técnicamente viable, sino también analíticamente fructífera: permitió ordenar el desempeño conjunto en Lengua y Matemática en perfiles distinguibles y, al integrarse con modelos explicativos, habilitó contrastar cómo se distribuyen los factores asociados dentro de cada perfil. Esta combinación metodológica constituye un avance respecto de análisis que separan estrictamente las áreas disciplinares, sin por ello abandonar el potencial del Machine Learning para modelar relaciones complejas en microdatos educativos.

7 Referencias

[1] Llach, J. J., & Cornejo, M. (2018). *Factores condicionantes de los aprendizajes en la escuela primaria y media.* In LIII Reunión Anual de la Asociación Argentina de Economía Política (La Plata, 14 al 16 de noviembre de 2018).

[2] Ministerio de Educación, Cultura, Ciencia y Tecnología. (2018) *Factores Condicionantes de los Aprendizajes. Primaria y Secundaria. Ciudad Autónoma de Buenos Aires*, Argentina. Recuperado de https://www.argentina.gob.ar/sites/default/files/factores_condicionantes_de_los_aprendizajes.pdf
[3] Llach, J. J., Schumacher, F., & Llach, J. (2006). *La segregación social en la educación primaria argentina*. Buenos Aires: Granica.
[4] Ministerio de Educación y Deportes. Secretaría de Evaluación Educativa (2016) *Aprender 20216. Informe de Resultados*. Recuperado de https://www.argentina.gob.ar/sites/default/files/reporte_nacional.pdf
[5] Villardón, J. L. V. (2007). Introducción al análisis de clúster. Departamento de Estadística, Universidad de Salamanca. 22p..
[6] Al-Omary, A. Y., & Jamil, M. S. (2006). A new approach of clustering based machine-learning algorithm. Knowledge-Based Systems, 19(4), 248-258.
[7] Suyal, M., & Sharma, S. (2024). A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning. Journal of Artificial Intelligence and Systems, 6(1), 85-95.
[8] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.
[9] Chicon, P. M. M., & Telocken, A. V. (2021). Otimização Aplicada a Técnica de Clusterização. Revista Interdisciplinar de Ensino, Pesquisa e Extensão, 9(1), 13-27.
[10] Juanita, S., & Cahyono, R. D. (2024). K-means clustering with comparison of elbow and silhouette methods for medicines clustering based on user reviews. Jurnal Teknik Informatika (JUTIF), 5(1), 283-289.