Informe Estandarización Perú Escala INDI, Parte 6: Análisis de Conglomerados

Muestra Nivel 4-5

Martín Vargas Estrada

2025-02-13 10:40:44.1605

Índice

Introducción

En este caso, procederemos a realizar un análisis de conglomerados de los datos, también llamado Cluster Analysis.

Marco Conceptual

Finalidad

Antes de pasar a los resultados de los análisis en sí, es necesario tener claro en qué consiste el Análisis de Conglomerados.

En general, hasta el momento nos hemos centrado en las variables (de modo principal en las Escalas del INDI, es decir, en sus respectivos puntajes), así como en las relaciones que hemos podido descubrir entre ellas. En esta última sección del Informe, vamos a enfocarnos más bien en los participantes, e intentaremos descubrir si es posible establecer agrupaciones entre tales participantes basándonos en las características cuya información hemos recolectado durante el proceso de evaluación.

Por poner un ejemplo, mediante el análisis de conglomerados podríamos llegar a establecer la existencia de un grupo de participantes que tienen en común su origen (por ejemplo, Lima), el nivel educativo materno (por ejemplo, Postgraduado), un mayor puntaje en la Escala C del INDI, etc., mientras que un segundo conglomerado podría estar formado por participantes cuyo origen es Cusco, tienen madres de nivel educativo Secundario completo, y un puntaje alto en la escala Socioemocional del INDI.

Estrategias

Para llegar a cumplir nuestra meta de establecer agrupamientos de participantes afines tenemos varias estrategias. No entraremos en el detalle técnico aquí, bastará mencionar que se ha escogido un método o modalidad de Análisis que nos permite incluir tanto variables numéricas (como Edad y los ya mencionados puntajes del INDI) como variables categóricas. Para determinar exactamente cuáles de estas últimas, hemos decidido usar como criterio los resultados del análisis de regresión. Ello nos permite estar seguros de que las variables consideradas tienen de antemano un impacto demostrado en las escalas INDI, evitando modelos con demasiados elementos, lo cual probablemente generaría modelos inútilmente complejos que no añadirían poder explicativo.

Parámetros de Análisis

La metodología elegida requiere que definamos de antemano la cantidad de conglomerados para generar nuestro modelo, dándonos la posibilidad de generar múltiple soluciones.

Por lo tanto, vamos a requerir algún criterio para establecer, si bien no necesariamente una solución "correcta" —ya que en estos casos no podemos establecer ese tipo de certidumbre, por la índole misma del análisis—, sí una solución más eficiente que las otras, tomando como criterio de "eficiencia" el mayor poder explicativo posible con el mínimo número posible de variables. En otras palabras, trataremos de encontrar el modelo más parsimonioso, que nos permita obtener los conglomerados más interpretables, a fin de no complejizar inútilmente el modelo, como ya se mencionó líneas arriba.

Para nuestro caso, ese parámetro de evaluación será el llamado "Coeficiente de Silueta". El coeficiente de silueta es una medida utilizada en análisis de conglomerado para evaluar la calidad de la agrupación de los datos. Se basa en la comparación de la cohesión dentro de un conglomerado con la separación respecto a los otros conglomerados.

- Un clustering o conjunto de conglomerados con un coeficiente de silueta promedio alto (cercano a 1) indica que los clusters están bien definidos.
- Valores cercanos a 0 sugieren clusters solapados.
- Valores negativos indican una mala asignación de puntos a clusters.

En este documento, para mayor legibilidad, hemos incluido solamente la solución que combina los coeficientes de silueta más altos, manteniendo a la vez la simplicidad necesaria para interpretar los clusters resultantes. Consideramos que una solución con muchos clusters, aun si se obtienen coeficientes ligeramente superiores, resultará mucho más difícil de interpretar, y por lo tanto menos útil, que soluciones con coeficiente solo ligeramente inferior pero de un número más manejable de clusters.

Asimismo, los resultados incluyen un gráfico que muestra la proporción de casos que "rebasan" respectivamente el cluster encontrado, así como otro que muestra la representación espacial de los clusters. Con esta información en mente, podemos pasar a revisar los resultados obtenidos.

Análisis de Conglomerados

A continuación, pasaremos a realizar el análisis de conglomerados, usando las variables siguientes:

1. Escala Cognitiva
2. Escala Motora
3. Escala Socioemocional
4. Escala Disposicional
5. Edad en Meses
6. Región
7. Modalidad
8. Departamento
9. Quintil
10. Inst. Materna

He aquí los resultados y su interpretación.

Resultados

Coeficiente de Silueta Promedio: 0.468 Número de Conglomerados: 2

Tabla de Coeficientes de Silueta por Clúster:

cluster	n	mean_sil_width
1	1322	0.48
2	874	0.44

Resumen de Variables Numéricas por Clúster:

Cluster	Escala_Cog.	Escala_Mot.	Escala_Soc	Escala_Dis.	Edad_Mes
1	70.90	22.53	60.88	21.72	60.83
2	116.18	30.14	70.31	28.38	64.80

Resumen de Variables Categóricas por Clúster:

Cluster	Región	Modalidad	Departamento	Quintil	Inst_Mat.
1	Costa	Jardín	Piura	4	Secundaria completo
2	Costa	Jardín	Lima Met.	4	Secundaria completo

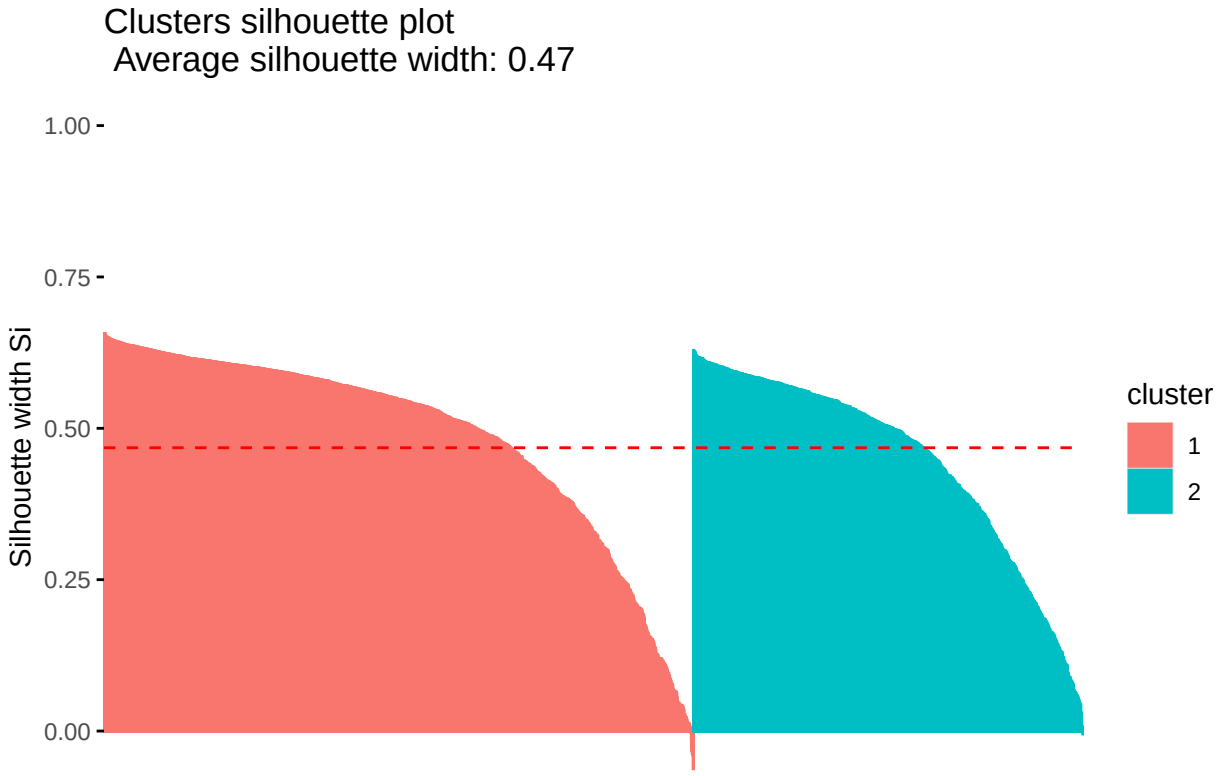
Clusters silhouette plot
Average silhouette width: 0.47
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2


Gráfico de Clústeres
Dim2 (15.8%)
2
0
−2
−4
−2.5
0.0
2.5
5.0
Dim1 (31.4%)
cluster
1
2

Interpretación

Análisis General

El algoritmo ha creado dos clusters a partir de los datos

El Cluster 1 es más numeroso que el Cluster 2, por poco menos de 500 casos (448, para ser exactos).

El Coeficiente de Silueta Promedio es 0.468 - Cluster 1: 0.48 - Cluster 2: 0.44

Generalmente, valores entre 0.25 y 0.50 indican una estructura de cluster moderadamente fuerte.

Sin embargo, el coeficiente de silueta no es muy alto, lo cual indica cierto solapamiento entre clusters, es decir los grupos no están perfectamente separados, como apreciamos en el Gráfico correspondiente.

Análisis de Variables Numéricas

Este análisis nos dice cómo se diferencian los dos grupos en las variables numéricas.

- El Cluster 2 tiene puntajes significativamente más altos en todas las escalas cognitivas, motivacionales y sociales.
- El Cluster 1 tiene puntajes más bajos en todas las variables, lo que implica que los individuos de este grupo pueden tener menor desempeño en la prueba.
- Edad en meses:
  - Cluster 1: 60.83 meses (~5 años y 1 mes)
  - Cluster 2: 64.80 meses (~5 años y 5 meses)
  - Diferencia: Cluster 2 es, en promedio, 4 meses mayor que Cluster 1.
  - Esto podría indicar, como ya hemos visto en niveles previos de análisis, que la edad influye en el desempeño de la prueba.

Análisis de Variables Categóricas

Este análisis nos ayuda a ver dónde viven estos participantes, su nivel socioeconómico y el nivel educativo de sus madres.

Los participantes de ambos clusters están más frecuentemente ubicados en la Costa y en la modalidad "Jardín" (educación inicial).

En el Cluster 1 hay mayoría de niños de Piura, mientras que el Cluster 2 tiene mayoría de Lima Metropolitana.

El nivel socioeconómico más frecuente similar (Quintil 4) en ambos clusters.

El nivel educativo materno más frecuente en ambos grupos es "Secundaria completa", lo que sugiere que esta variable no parece ser un factor diferenciador en el clustering.

Conclusiones Generales

1. Diferencia en desempeño del INDI:
   - El Cluster 2 tiene un desempeño superior en todas las escalas.
   - Esto puede estar relacionado con factores como edad, ubicación geográfica o factores no medidos asociados (calidad educativa, acceso a recursos, etc.).
2. Ubicación geográfica:
   - Piura (Cluster 1) vs. Lima Metropolitana (Cluster 2).
   - Esto sugiere que los niños de Lima pueden estar obteniendo mejores resultados en la prueba.
   - Puede haber diferencias en la calidad educativa o en el acceso a oportunidades de aprendizaje.
3. Edad como factor de influencia:

- El Cluster 2 es un poco mayor (4 meses mayor, en promedio).
- Aunque parece una diferencia pequeña, a esta edad (5 años), meses adicionales pueden representar una ventaja en términos de maduración, desarrollo cognitivo y habilidades sociales.