Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Histogramas, Kernels y Métodos No Supervisados usando la EPH

Trabajo Práctico N°2 - Grupo 15

DOCENTE: NOELIA ROMERO

ASIGNATURA: TALLER DE PROGRAMACIÓN

Repositorio GitHub - Grupo 15

INTEGRANTES: JULIÁN DELGADILLO MARÍN, ALEJANDRO ALCOCER, CHRISTIAN CAMPOS
POSGRADO: MAESTRÍA EN ECONOMÍA APLICADA

Fecha de Entrega: Viernes 24 de Octubre de 2025

Resumen

El trabajo aplica herramientas de análisis descriptivo y métodos no supervisados sobre la Encuesta Permanente de Hogares (EPH) para los años 2005 y 2025. Se construyen variables homogéneas de edad, educación, ingreso y horas trabajadas, y se analizan sus distribuciones mediante histogramas y kernels. Posteriormente, se implementan técnicas de reducción de dimensionalidad (PCA) y agrupamiento (k-medias y jerárquico) para identificar patrones socioeconómicos entre hogares pobres y no pobres. Los resultados muestran diferencias consistentes en la estructura educativa, el ingreso y el trabajo, confirmando la utilidad de los métodos no supervisados para caracterizar heterogeneidades sociales.

Palabras clave: EPH; histogramas; kernel; PCA; clustering; pobreza.

1. Análisis

[

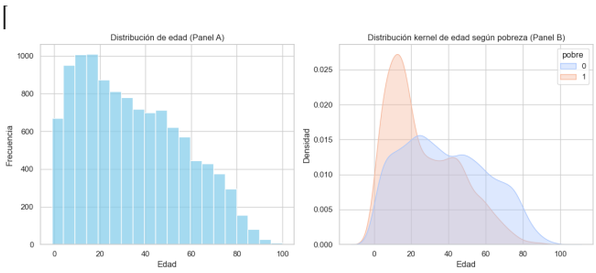

Figura 1. Distribución de la edad en la EPH (Panel A: histograma; Panel B: densidad kernel por condición de pobreza).

El análisis de la variable *edad* (Figura 1) caracteriza la estructura demográfica de la población mediante dos enfoques: un histograma de la distribución general (Panel A) y una densidad Kernel segmentada por condición de pobreza (Panel B). El Panel A muestra una distribución sesgada a la derecha, concentrada en edades tempranas, con un pico alrededor de los 10 años y disminución progresiva hacia edades mayores, típica de una población joven. El Panel B evidencia diferencias entre hogares: los pobres concentran mayor densidad en edades entre 10 y 20 años, indicando alta dependencia infantil, mientras que los no pobres presentan una distribución más aplanada y desplazada hacia edades medias y avanzadas. En conjunto, la superposición de distribuciones confirma que la pobreza se asocia a una población más joven, con implicaciones sobre variables del mercado laboral, como ingreso y horas trabajadas, que se analizarán posteriormente.

Variable	N	Media	Desv. Std.	Mín.	P25	P50	P75	Máx.
educ (años)	11,352	7.77	4.87	0.00	3.00	6.00	12.00	16.00

Cuadro 1. Estadísticos descriptivos de la variable *educ* (años de educación formal).

El análisis descriptivo de la variable *años de educación formal* evidencia una alta heterogeneidad en el nivel educativo de la población. El promedio alcanza los 7.8 años, acompañado de una desviación estándar de 4.87, lo cual revela una amplia dispersión en la distribución. La mediana es de 6 años, indicando que al menos la mitad de las personas no supera la educación primaria completa. Además, el 25 % cuenta con 3 años o menos de estudio, mientras que el 75 % no supera los 12 años, valor asociado a la finalización de la educación secundaria. La amplitud entre los valores mínimos (0) y máximos (16) confirma una estructura educativa polarizada y marcada por brechas significativas en acceso y culminación escolar.

Condición de pobreza	Frecuencia	Porcentaje (%)
No pobres (0)	7,808	68.8
Pobres (1)	3,544	31.2

Años de educación (*educ*)	Probabilidad promedio de pobreza
0	0.436
3	0.399
6	0.362
9	0.344
12	0.182
15	0.110
16	0.199

Cuadro 2. Distribución general de hogares según condición de pobreza y probabilidad promedio de pobreza por nivel educativo.

El Cuadro 2 evidencia que el 31.2 % de los hogares se encuentran en situación de pobreza, estableciendo la prevalencia que el análisis busca explicar. Asimismo, se identifica una relación inversa marcada entre los años de educación y la probabilidad de ser pobre: la ausencia de escolaridad se asocia a un riesgo cercano al 44 %, mientras que la finalización del nivel secundario reduce dicho riesgo a aproximadamente 18 %. Los niveles educativos superiores presentan las menores probabilidades de pobreza, lo que confirma el efecto protector del capital educativo y resalta la relevancia de garantizar acceso y culminación escolar para mejorar las condiciones socioeconómicas de la población.

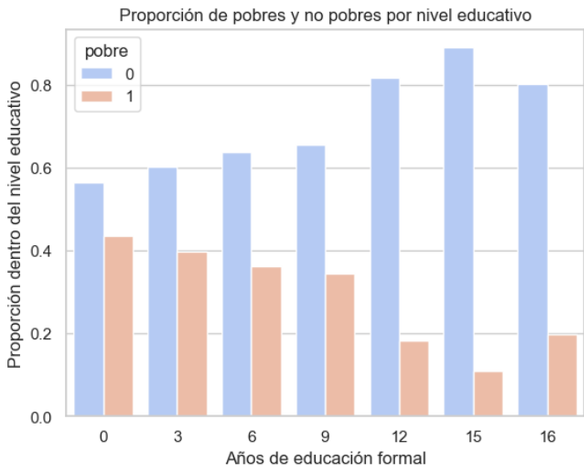

Figura 2. Proporción de hogares pobres y no pobres según años de educación formal en la EPH (2005–2025).

El Gráfico 2 confirma visualmente la relación inversa entre educación y pobreza previamente identificada en el Cuadro 2. En los niveles educativos bajos (0–9 años), la proporción de hogares pobres oscila entre 35 % y 43 %, reflejando la limitada capacidad protectora de la educación primaria e incompleta. El punto de inflexión se observa en los 12 años de estudio, correspondientes a la finalización de la educación secundaria, donde la proporción de pobreza desciende a aproximadamente 18 % y la mayoría de los individuos se ubica entre los no pobres. Los niveles superiores (15–16 años) muestran la menor incidencia de pobreza, con valores cercanos al 10 %, lo que reafirma el papel de la educación superior como principal mecanismo de protección frente a la vulnerabilidad económica.

Año	N	Media	Desv. Std.	Mín.	P25	P50	P75	Máx.
2005	7,043	416,168.05	550,615.47	3,093.0	185,580.0	309,300.0	516,221.7	16,099,065.0
2025	4,309	1,816,290.22	1,947,736.42	40,000.0	751,000.0	1,300,000.0	2,200,000.0	20,180,000.0

Cuadro 3. Estadísticos descriptivos del ingreso total familiar (ajustado a precios de 2025) por año de observación.

El análisis descriptivo del ingreso para los años 2005 y 2025 evidencia un incremento nominal considerable, coherente con un contexto de alta inflación y crecimiento económico. En ambos casos, la media es sustancialmente mayor que la mediana, lo que confirma una distribución sesgada a la derecha. Asimismo, el notable aumento del rango intercuartílico y de la desviación estándar revela una mayor dispersión en la parte central y alta de la distribución. En conjunto, estos resultados sugieren que, pese al crecimiento generalizado de los ingresos, la desigualdad se ha mantenido elevada e incluso podría haberse intensificado, destacando el ingreso como un factor central para caracterizar heterogeneidades socioeconómicas en los análisis posteriores.

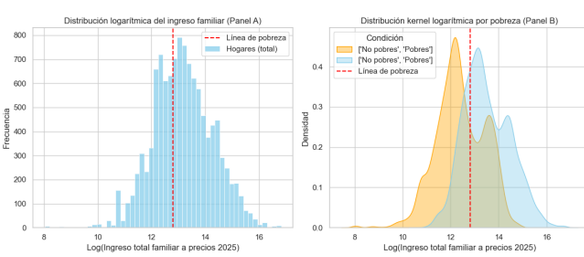

Figura 3. Distribución logarítmica del ingreso total familiar a precios de 2025 (Panel A: histograma total; Panel B: densidad kernel por condición de pobreza, con línea de pobreza indicada).

La Figura 3 muestra la distribución logarítmica del ingreso familiar (precios 2025) con la Línea de Pobreza (roja discontinua). La transformación logarítmica reduce la asimetría, concentrando la mayoría de los hogares entre 12 y 15, con un pico alrededor de 13. Aproximadamente el 31.2 % se ubica por debajo del umbral. La estimación Kernel por condición de pobreza evidencia hogares pobres concentrados en ingresos bajos y hogares no pobres dispersos y bimodales, revelando heterogeneidad interna y justificando el uso de PCA y k-medias para identificar subgrupos socioeconómicos.

Método de estimación	Descripción	Tasa de pobreza (%)
Variable oficial (pobre)	Ajustada por tamaño y composición del hogar (líneas específicas por equivalencia)	31.22
Línea única CBT	Ingreso total familiar menor a $365,177 (CBT 2025)	38.41

Cuadro 4. Comparación de tasas de pobreza estimadas según dos metodologías.

El Cuadro 4 muestra que la tasa de pobreza es de 31.22 % usando la metodología oficial ajustada por tamaño y composición del hogar, mientras que una Línea Única CBT eleva la tasa a 38.41 %. Esta diferencia de 7.19 puntos evidencia la alta sensibilidad del indicador al método y confirma que el ajuste por equivalencia proporciona una estimación más precisa. Por ello, el estudio utiliza la Variable Oficial para clasificar hogares pobres y no pobres, asegurando que los patrones socioeconómicos identificados por los métodos no supervisados se basen en una medida confiable de pobreza.

Variable	N	Media	Desv. Std.	Mín.	P50	Máx.
horastrab (horas/semana)	3,661	26.03	25.26	0.00	24.00	99.00

Cuadro 5. Estadísticos descriptivos de las horas trabajadas por el jefe/a de hogar.

Indicador	2005	2025	Total
Cantidad de observaciones	7,043	4,309	11,352
Observaciones con NA en `pobre`	0	0	0
Cantidad de pobres	2,210	1,334	3,544
Cantidad de no pobres	4,833	2,975	7,808
Cantidad de variables	186	186	186

Cuadro 6. Resumen general de la base de datos por año.

El Cuadro 6 resume la estructura de la base de datos por año y la distribución de la variable de pobreza. La muestra totaliza 11,352 observaciones con 186 variables, aunque se reduce de 7,043 en 2005 a 4,309 en 2025, lo que puede afectar la precisión de las estimaciones recientes. No existen datos faltantes en la variable pobre, asegurando integridad analítica. La tasa de pobreza se mantiene estable: 31.38 % en 2005 y 30.96 % en 2025, resultando en un 31.22 % para el total de la muestra. Esta estabilidad, a pesar del fuerte crecimiento nominal del ingreso, indica que los factores estructurales que determinan la pobreza (educación, empleo e ingresos reales) persisten, justificando la necesidad de aplicar métodos no supervisados para identificar las características subyacentes que mantienen a un tercio de la población en situación de pobreza.

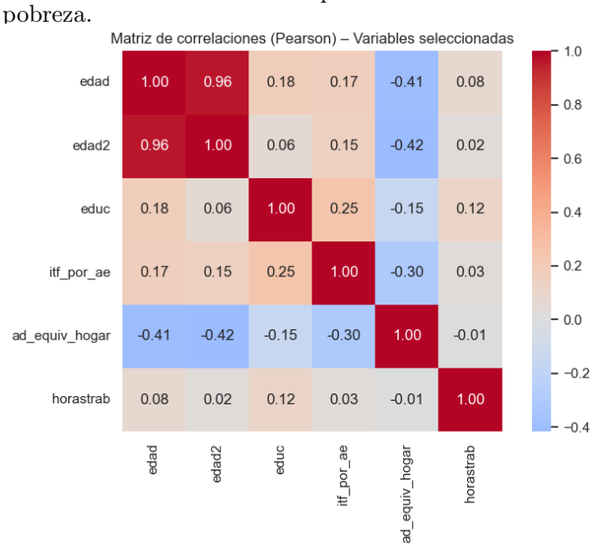

Figura 4. Matriz de correlaciones (coeficiente de Pearson) entre las variables utilizadas para el análisis no supervisado: edad, edad^2, educación, ingreso total familiar, adultos equivalentes por hogar y horas trabajadas.

La Figura 4 muestra que Edad–Edad^2 está altamente correlacionada (0.96), justificando PCA. Educación e Ingreso se correlacionan moderadamente (0.25), mientras

que hogares grandes y jóvenes presentan menor ingreso (-0.30 y -0.41). Horas trabajadas tienen poca correlación. Ingreso per cápita, Adultos Equivalentes y Educación se destacan como ejes principales para métodos no supervisados.

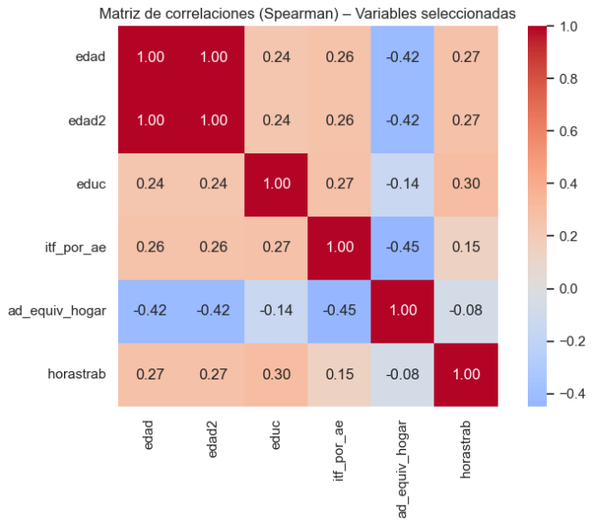

Figura 5. Matriz de correlaciones (coeficiente de Spearman) entre las variables del análisis no supervisado: edad, edad^2, educación, ingreso total familiar, adultos equivalentes por hogar y horas trabajadas.

La Figura 5 muestra que el ingreso per cápita ajustado se correlaciona negativamente con el tamaño del hogar ($\rho = -0{,}45$) y positivamente con educación ($\rho = 0{,}27$). Edad y tamaño del hogar se correlacionan negativamente ($\rho = -0{,}42$) y educación y horas trabajadas positivamente ($\rho = 0{,}30$). Ingreso, tamaño del hogar y educación constituyen los ejes principales para PCA y clustering.

Componente	Varianza Explicada (%)	Varianza Acumulada (%)
PC1	39.71	39.71
PC2	19.99	59.70
PC3	16.75	76.46
PC4	12.79	89.25
PC5	10.26	99.51
PC6	0.49	100.00

Cuadro 7. Proporción de varianza explicada por los componentes principales del análisis PCA.

El Cuadro 7 muestra que PC1-PC3 capturan 76.5 % de la varianza y representan los ejes principales de diferenciación socioeconómica; incluir PC4 alcanza 90 %. Tres o cuatro componentes permiten reducir la dimensionalidad sin pérdida significativa, facilitando el clustering.

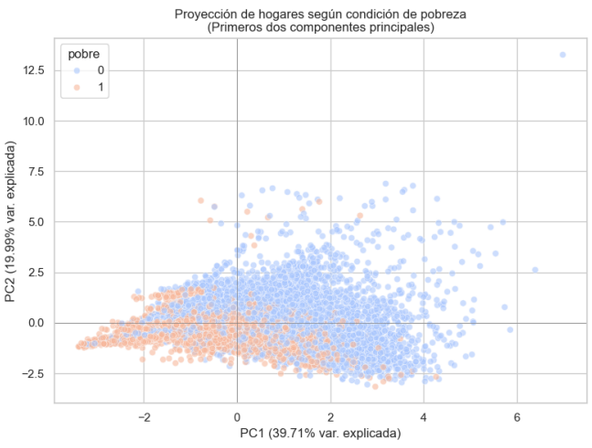

Figura 6. Proyección de los hogares en los dos primeros componentes principales (PCA), diferenciados por condición de pobreza.

La Figura 6 proyecta los hogares en PC1-PC2. PC1 (39.7 %) separa claramente Pobres (negativos) de No Pobres (positivos), mientras que PC2 (20 %) refleja heterogeneidad demográfica. El solapamiento cercano al origen indica hogares vulnerables, mostrando que el clustering permite identificar subgrupos dentro de los no pobres.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
edad	0.596	-0.247	0.152	0.178	0.177	0.706
edad2	0.583	-0.337	0.097	0.089	0.190	-0.702
educ	0.203	0.640	0.033	0.714	-0.174	-0.091
itf_por_ae	0.269	0.545	-0.385	-0.408	0.562	0.009
ad_equiv_hogar	-0.433	-0.107	0.248	0.392	0.766	-0.016
horastrab	0.066	0.328	0.870	-0.360	-0.015	-0.033

Cuadro 8. Cargas factoriales (loadings) de las variables originales sobre los seis componentes principales.

El cuadro 8 muestra las cargas de los Componentes Principales. PC1 (39.7 %) refleja estabilidad socioeconómica y edad; PC2 (20 %) capta capital humano e ingreso; PC3 y PC4 destacan horas trabajadas y combinaciones de educación e ingreso. Los PC diferencian hogares pobres (bajo PC1) de no pobres (alto PC1) y explican 60 % de la varianza, sirviendo de base para el clustering.

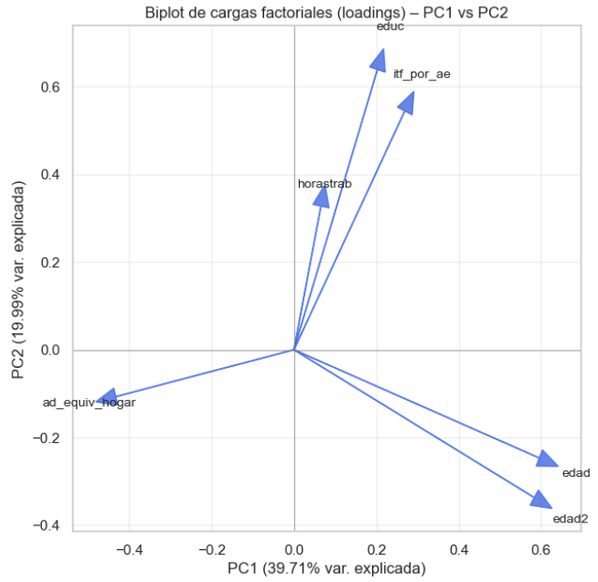

Figura 7. Biplot de cargas factoriales (loadings) de las variables en los dos primeros componentes principales del PCA.

La Figura 7 muestra el biplot PC1-PC2. PC1 (39.7 %) refleja edad y tamaño del hogar, discriminando pobreza estructural; PC2 (20 %) capta educación e ingreso. El gráfico evidencia subgrupos socioeconómicos, justificando el clustering más allá de la clasificación binaria.

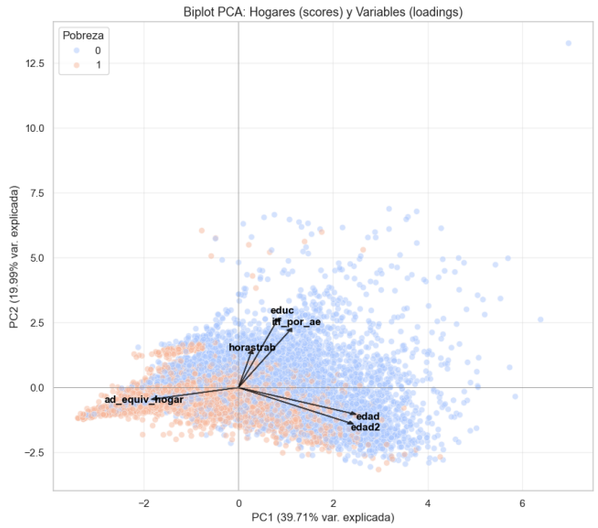

Figura 8. Biplot del análisis de componentes principales (PCA) que muestra simultáneamente los hogares (scores) y las variables (loadings), diferenciados por condición de pobreza.

La Figura 8 muestra el biplot PC1-PC2. PC1 (39.7 %)

refleja estabilidad socioeconómica y edad, separando hogares pobres (jóvenes, dependientes) de no pobres (adultos, baja dependencia). PC2 (20 %) capta educación e ingreso. La zona central evidencia limitaciones de la clasificación binaria, mientras que clustering identifica subgrupos como "pobre por estructuraz "no pobre vulnerable". Las variables clave (edad, ad_equiv_hogar, educ, itf_por_ae) explican la diferenciación socioeconómica.

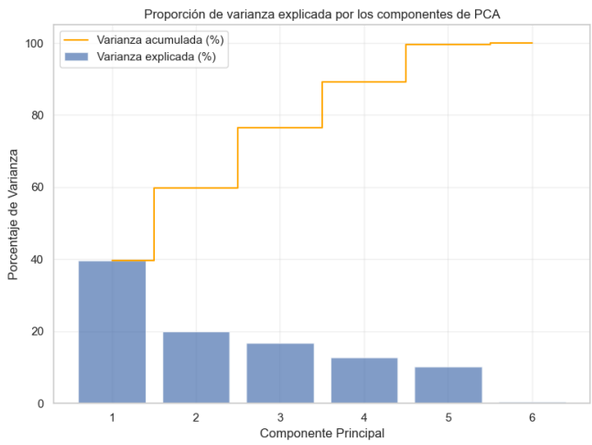

Figura 9. Proporción de varianza explicada y acumulada por los componentes principales del análisis PCA.

La Figura 9 muestra el gráfico de codo del PCA. PC1-PC3 explican 76.5 % de la varianza y PC4 lleva cerca del 90 %. El "codo" indica que 3-4 componentes capturan la mayor parte de la heterogeneidad, validando la reducción de dimensionalidad para clustering.

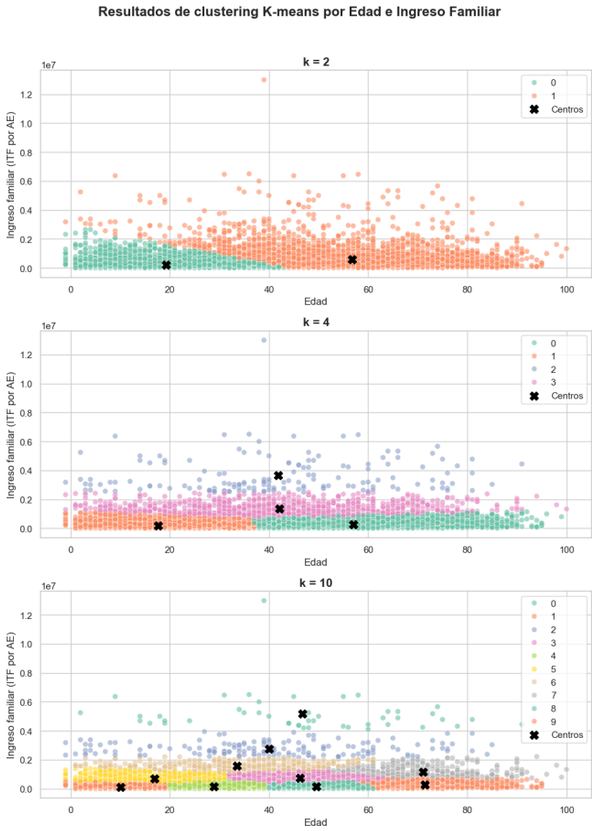

Figura 10. Resultados del algoritmo k-medias aplicados a la relación entre edad e ingreso familiar per cápita equivalente, para distintos números de clusters ($k = 2$, $k = 4$ y $k = 10$).

La Figura 10 muestra el clustering k-medias sobre Edad e Ingreso per cápita. $k = 2$ reproduce la dicotomía pobre/no pobre; $k = 4$ identifica clústeres interpretables (pobreza extrema, vulnerables, clase trabajadora, clase media/alta), alineados con PC1-PC2. $k = 10$ evidencia sobre-ajuste. Se concluye que $k = 4$ equilibra separación y utilidad para caracterizar patrones socioeconómicos.

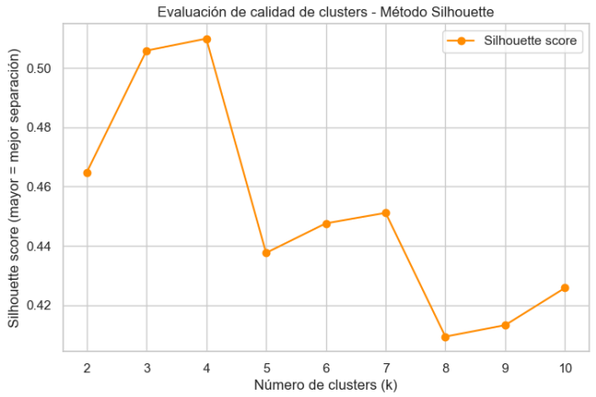

Figura 11. Evaluación de la calidad de los clusters mediante el método del coeficiente Silhouette para distintos valores de k.

La Figura 11 muestra el Coeficiente Silhouette para distintos k. El máximo se alcanza en $k = 4$ (0.52), indicando la mejor separación de clústeres. $k < 4$ oculta heterogeneidades y $k > 4$ genera sobre-ajuste, validando $k = 4$ para segmentar patrones socioeconómicos más allá de la clasificación binaria.

Cluster	**Pobres (%)**	**No pobres (%)**
0	48.91	51.09
1	21.95	78.05
Pureza aproximada	64.57 %	

Cuadro 9. Distribución porcentual de pobreza y pureza del modelo de clustering con $k = 2$.

Cluster	**Edad** (media)	**Edad** (mediana)	**ITF por AE** (media)	**ITF por AE** (mediana)	**N** (obs.)
0	57.07	55.00	269,437	184,657	4027
1	17.64	17.00	196,449	122,253	5995
2	41.90	41.00	3,652,295	3,184,713	150
3	42.04	42.00	1,343,806	1,240,000	1067

Cuadro 10. Resumen descriptivo de los grupos formados mediante clustering k-means ($k = 4$).

Los Cuadros 9 y 10 muestran las estadísticas de los cuatro clústeres óptimos ($k = 4$) según Edad e ITF por AE. Clúster 1: hogares jóvenes, dependientes y bajos ingresos; Clúster 0: adultos mayores con ingresos bajos; Clúster 3: clase media estable; Clúster 2: élite de alto ingreso. Esto evidencia heterogeneidad dentro de pobres y no pobres, confirmando la utilidad de métodos no supervisados para identificar subgrupos socioeconómicos más finos.

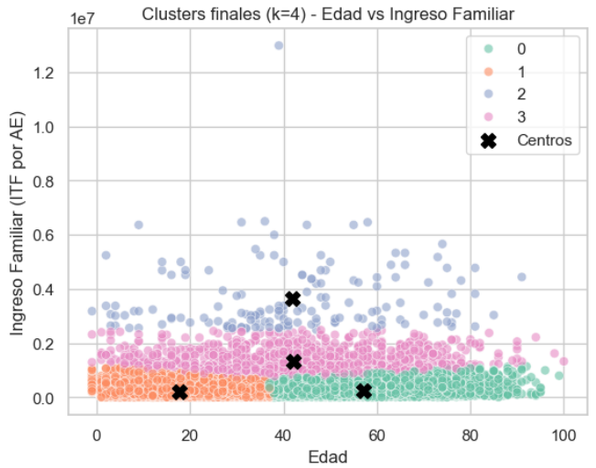

Figura 12. Segmentación final de hogares según edad e ingreso familiar per cápita equivalente utilizando el algoritmo k-medias con $k = 4$.

La Figura 12 confirma visualmente los cuatro patrones socioeconómicos identificados: Clúster 1 (Naranja, Po-

breza Joven; edad ≈ 17, ITF ≈ 50,000) agrupa hogares con alta dependencia infantil y bajos ingresos; Clúster 0 (Verde/Cian, Pobreza Mayor; edad ≈ 57, ITF ≈ 50,000) representa pobreza en adultos mayores; Clúster 3 (Rosa, Clase Media/Estable; edad ≈ 42, ITF ≈ 150,000) identifica hogares no pobres con ingresos estables; Clúster 2 (Azul/Morado, Élite; edad ≈ 41, ITF ≈ 400,000) concentra hogares de alto ingreso. La segmentación valida la influencia de PC1 (estructura demográfica) y PC2 (capital humano/ingreso), diferenciando dos tipos de pobreza y dos de no pobreza, confirmando la heterogeneidad social y la utilidad del clustering frente a la clasificación binaria.

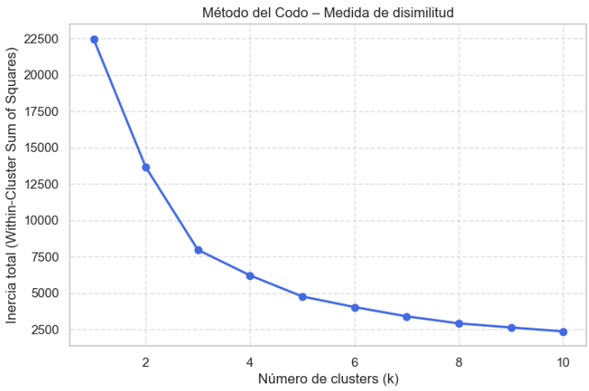

Figura 13. Determinación del número óptimo de clusters mediante el método del codo, utilizando la inercia total (Within-Cluster Sum of Squares) como medida de disimilitud.

La Figura 13 muestra la Inercia Total (WCSS) en función de k, indicando que la reducción de dispersión es significativa al pasar de $k = 2$ a $k = 3$ y $k = 4$, y se aplana a partir de $k = 4$–5. Este punto de inflexión, consistente con el máximo del Coeficiente Silhouette (Figura 11), confirma que $k = 4$ es el número óptimo de clústeres, equilibrando compacidad interna y simplicidad. La elección valida la segmentación final de cuatro patrones socioeconómicos (Patrones 0, 1, 2 y 3) para caracterizar la heterogeneidad social de manera robusta y fiable.

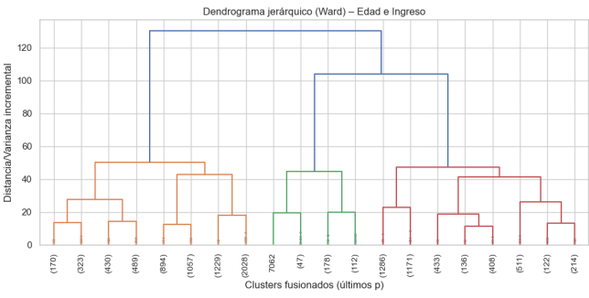

Figura 14. Dendrograma del análisis de clustering jerárquico utilizando el método de Ward sobre las variables edad e ingreso familiar per cápita equivalente.

La Figura 14 visualiza la fusión de clústeres según la disimilitud, indicando que los saltos más grandes ocurren entre $k = 2$ y $k = 3$, y luego a niveles que soportan la separación en cuatro grupos. Una inspección a un nivel de corte medio-bajo (altura ≈ 50) revela cuatro ramas principales: izquierda (Pobreza Joven y Mayor, Clústeres 0 y 1), derecha (Clase Media/Estable, Clúster 3), y central (Élite/Alto Ingreso, Clúster 2). El dendrograma valida la estructura encontrada con k-medias, mostrando subgrupos homogéneos dentro de cada patrón y confirmando que la elección de $k = 4$ refleja de manera robusta la heterogeneidad socioeconómica de los hogares.

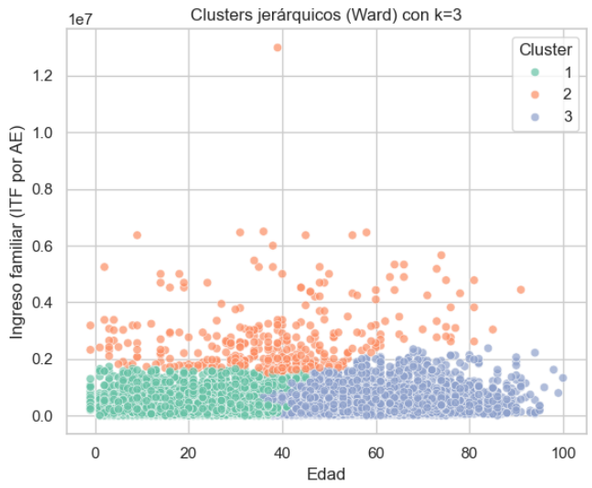

Figura 15. Segmentación de hogares obtenida mediante clustering jerárquico (método de Ward) con $k = 3$, en función de la edad y el ingreso familiar per cápita equivalente.

La Figura 15 muestra la segmentación jerárquica en tres clústeres: Clúster 1 (Clase Media/Élite, ingresos medios-altos), Clúster 2 (Pobreza Mayor, bajos ingresos, adultos) y Clúster 3 (Pobreza Joven, bajos ingresos, jóvenes). Esta división confirma las grandes categorías socioeconómicas, pero $k = 3$ no distingue Clase Media de Élite, validando $k = 4$ en k-medias para mayor granularidad.

Cluster	No pobres (0)	Pobres (1)
0	0.690	0.310
1	0.686	0.314

Cuadro 11. Distribución proporcional de pobres y no pobres por cluster ($k = 2$).

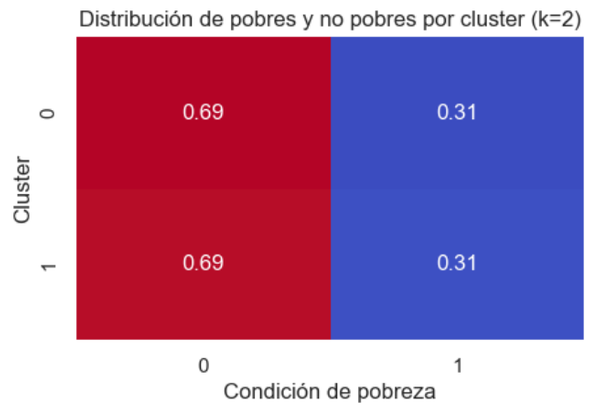

Figura 16. Distribución conjunta de hogares pobres y no pobres según los clusters obtenidos con el modelo k-medias ($k = 2$).

El Cuadro 11 muestra que ambos clústeres ($k = 2$) tienen composiciones similares: 31 % pobres y 69 % no pobres. La Figura 16 evidencia esta superposición, indicando que $k = 2$ no discrimina efectivamente pobreza y que se requieren al menos cuatro clústeres para capturar diferencias significativas entre pobres (jóvenes vs. mayores) y no pobres (clase media vs. élite).