



Lic. en Ciencias del Comportamiento

Trabajo Práctico N° 2

Alumnas

Agustina Kiessner

Isabela Nicola

Clara Mollón

Profesores

María Noelia Romero

Tomas Enrique Buscaglia

Ignacio Anchorena

Asignatura

Ciencia de datos - Tutorial 3

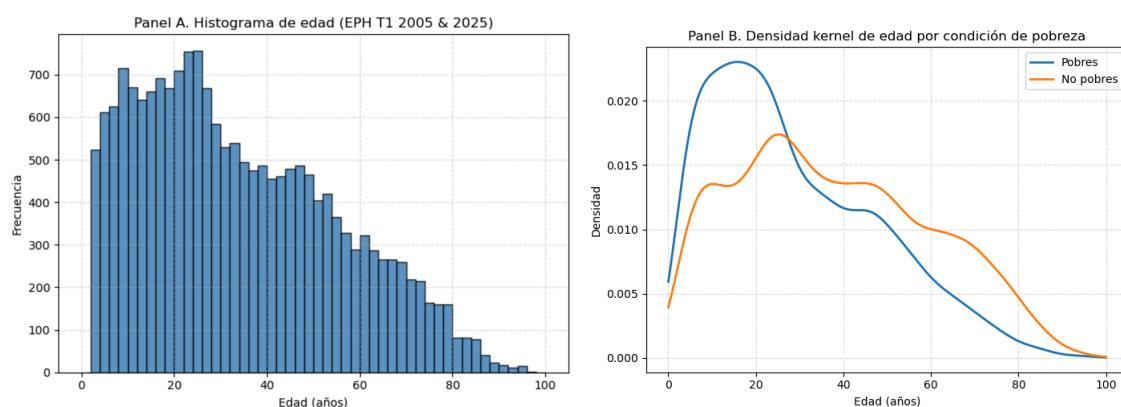
Fecha de presentación

26 de Septiembre de 2025

Parte I: Creación de variables, histogramas, kernels y resumen de la base de datos final

1. En el panel A (Figura 1) la distribución de edades presenta una mayor concentración en niños, adolescentes y jóvenes adultos, con un pico alrededor de los 20 años donde luego de este pico (30–40 años) la frecuencia de observaciones comienza a descender. Esto evidencia que la mayoría de los encuestados corresponde a población joven, con una proporción significativa de adultos mayores. En el panel B (Figura 1), la curva de pobreza está más concentrada entre los 15 a 20 años sugiriendo una incidencia de pobreza en la población joven. En la distribución de no pobres se puede ver que es más “aplanada” donde se desplaza hacia edades medias y adultas, con mayor densidad relativa a partir de los 30–40 años, indicando una segmentación etaria de la pobreza.

Figura 1



Nota. La figura izquierda es el Panel A, donde se presenta el histograma de la distribución de edad de los encuestados de la EPH en 2005 y 2025 en la región del NOA. A su derecha se presenta el Panel B se presenta distribución de kernels para los pobres y no pobres.

2. Se construyó la variable “educ” como años de educación formal acumulados integrando 3 variables: nivel alcanzado (CH12), finalización del nivel (CH13) y último año aprobado (CH14). Es importante señalar que en la variable CH12 el valor 9 corresponde a la categoría Educación especial, y que en la variable CH14 el valor 98 hace referencia al mismo tipo de educación. Dado que no resulta posible traducir esta trayectoria a un número de años de educación formal comparable con el resto de los niveles, ambas categorías fueron recodificadas como valores perdidos (NaN) para los fines de esta consigna. En la Tabla 1, la escolaridad en el NOA presenta una distribución amplia desde 0 a 19 años con mediana de 10 años y promedio parecido de 9.26, lo que sugiere asimetría hacia valores bajos, es decir que existe un segmento no trivial con muy pocos años de educación que tracciona el promedio hacia abajo. Al mismo tiempo, el máximo de 19 años indica presencia de trayectorias largas de educación, aunque menos frecuentes. En conjunto, el patrón es consistente con heterogeneidad educativa marcada en la región con una concentración en torno a la secundaria (≈ 10 – 12 años) y que hay un segmento importante con baja escolaridad que tira el promedio hacia abajo.

Tabla 1

Estadísticas de educ (años de educación)

Promedio	9.26 años
Desvío Estándar	4.63
Mínimo	0 años
Mediana (p50)	10 años
Máximo	19 años

3. La distribución del panel A (Figura 2) muestra una elevada concentración de personas en tramos bajos de ingreso y poca frecuencia en valores altos donde la línea de pobreza queda ubicada hacia el extremo izquierdo del soporte, indicando que una fracción relevante de hogares se sitúa por debajo del umbral aún tras homogeneizar el poder de compra. Además, se observa que la mayoría de los ingresos familiares se agrupan en un rango bastante reducido en comparación con la amplitud total de la distribución. En la Figura 3 se presenta el Panel B donde la curva de pobres se concentra bien a la izquierda de la línea, con menor dispersión y picos en niveles bajos de ingreso; la de no pobres aparece desplazada a la derecha, con mayor dispersión y cola extensa. El solapamiento es acotado y ocurre cerca del umbral, coherente con una transición nítida entre ambos grupos una vez llevados todos los montos a precios de 2025, por lo cual esto sugiere que la condición de pobreza se asocia con una clara segmentación de ingresos, donde la movilidad hacia el grupo de no pobres requiere superar un umbral relativamente definido. Es importante destacar que para la línea de pobreza utilizamos la canasta básica de 2025 ya que los precios están actualizados a los de ese año.

Figura 2

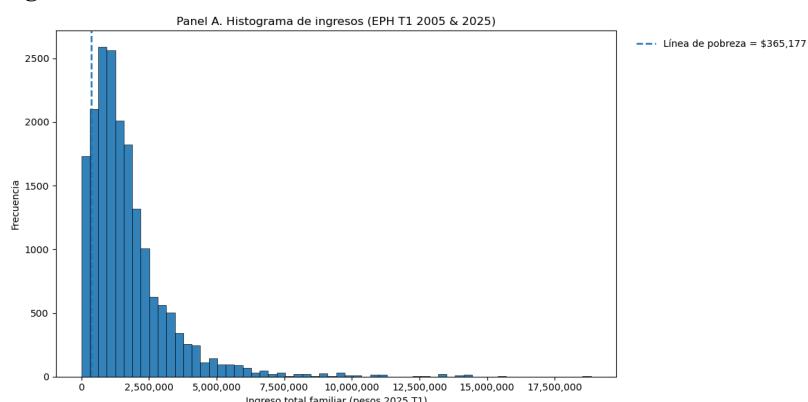
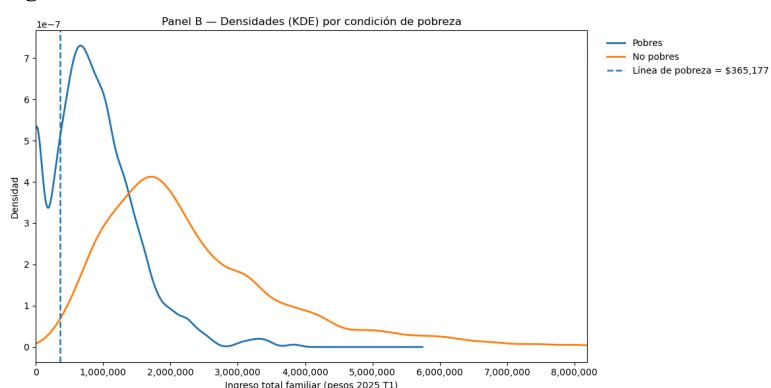


Figura 3



4. Luego de construir la variable de “horastrab” sumando las horas de la principal ocupación y otras ocupaciones solo para los jefes de hogar. Para esta variable se realizaron estadísticas descriptivas que se presentan en el Anexo 1. Como podemos ver, el promedio de horas trabajadas resulta 29,9 horas semanales, con un desvío estándar de 24,2. El valor mínimo observado es 0 horas (personas que no trabajaron en la semana de referencia), mientras que el máximo alcanza 126 horas, reflejando situaciones de pluriempleo e intensas cargas laborales. La mediana se ubica en 30 horas, muy próxima al promedio, lo que sugiere cierta simetría en la distribución, aunque con alta dispersión y presencia de casos extremos. En total, la variable fue calculada sobre 4.236 observaciones.

5.

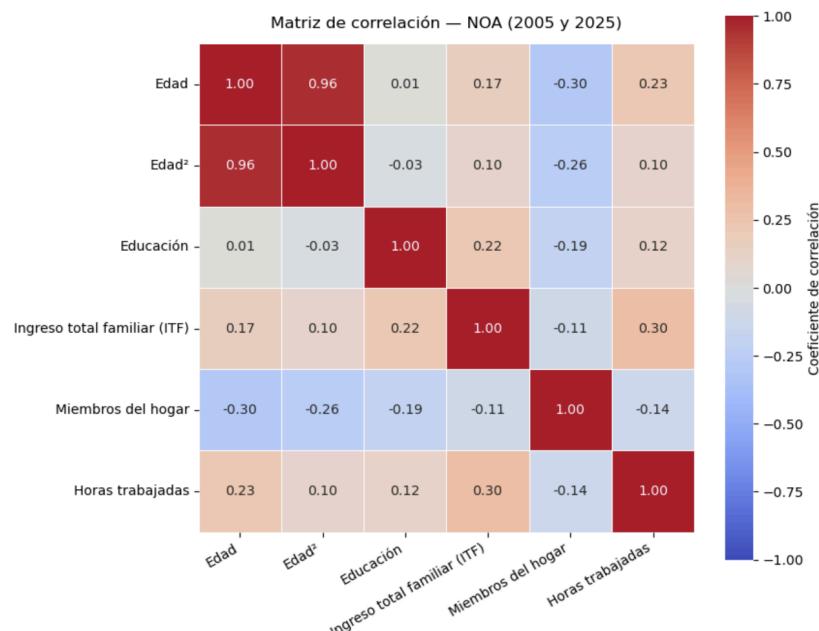
Tabla 2. Resumen de la base final para la región NOA

	2005	2025	Total
Cantidad observaciones	9.009	9.615	18.624
Cantidad de observaciones con NAs “Pobre”	0	0	0
Cantidad de Pobres	4.088	4.917	9.005
Cantidad de No Pobres	4.921	4.698	9.619
Cantidad de variables limpias y homogeneizadas	18	18	18

La tabla 2 presenta el tamaño muestral y el estado de pobreza en 2005 y 2025, 9.009 y 9.615 observaciones (18.624 total), sin datos faltantes. Con la variable creada anteriormente, se registran 4.088 pobres y 4.921 no pobres en 2005, 4.917 y 4.698 en 2025 (totales: 9.005 y 9.619). La incidencia sube de 45,4% a 51,1% (48,4% total). Se encontraron 18 variables limpias y homogeneizadas en ambos años, asegurando comparabilidad.

Parte 2

1. Figura 4

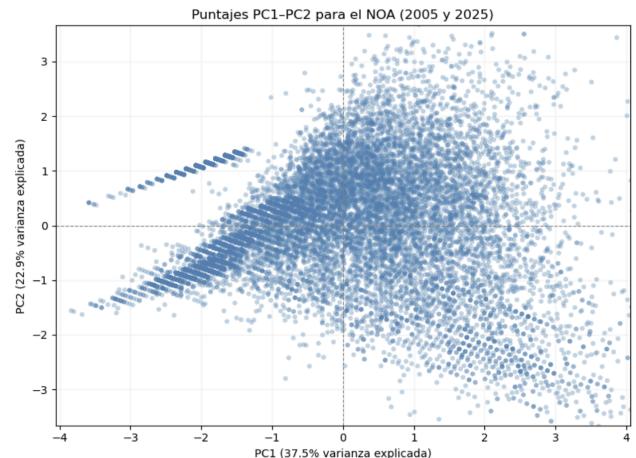


Calculamos la matriz de correlaciones entre seis predictores (edad, edad², educación, ingreso total familiar, miembros del hogar y horas trabajadas) con casos completos ($N=13.362$) de 2005 y 2025. La variable horas trabajadas se definió como la suma de PP3E_TOT y PP3F_TOT (se considera faltante si ambas lo estaban), sin restringir a jefes y jefas de hogar para maximizar el tamaño muestral. Los resultados muestran una muy alta correlación entre edad y edad² (0,96), esperable por ser una transformación no lineal, también muestra cómo ITF correlaciona positivamente con educación (0,22) y horas trabajadas (0,30), y levemente con edad (0,17). Asimismo, miembros del hogar se asocia negativamente con edad (-0,30) y con ITF (-0,11). Del mismo modo, la relación entre edad y horas trabajadas es baja y positiva (0,23), la educación se correlaciona negativamente con miembros del hogar (-0,19) y débilmente positivo con horas trabajadas (0,12). En general, las asociaciones son bajas a moderadas, salvo la redundancia entre edad y edad².

2.

Como se puede observar en la Figura 5, aplicamos PCA sobre las seis variables estandarizadas y graficamos los puntajes de los dos primeros componentes. El PC1 explica un 37,5% de la varianza y el PC2 un 22,9%, por lo que la varianza acumulada es de un 60,3%, con un $N=13.362$ ya que usamos los casos completos sin valores faltantes. La nube se alarga sobre el eje de PC1, mostrando que ese componente concentra la mayor parte de las diferencias entre individuos, el PC2 aporta variación adicional pero menor. La alta concentración de puntos alrededor del eje (0,0) indica que muchos casos están cerca de los valores promedio en ambas direcciones. En valores altos de PC1 se observa más dispersión.

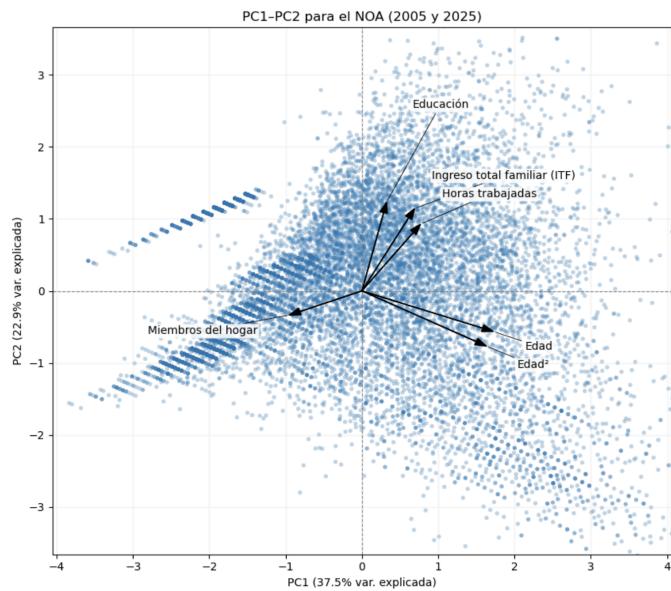
Figura 5



3. El gráfico muestra los *loadings* de las seis variables estandarizadas sobre los dos primeros componentes. En PC1 destacan edad (0,62) y edad² (0,59) con peso positivo, también pesan las horas trabajadas (0,28) y el ingreso total familiar (0,25), mientras que miembros del hogar recibe un peso negativo (-0,34). Por lo tanto, el PC1 aumenta con mayor edad, más horas e ingreso, y disminuye a medida que el hogar es más numeroso.

En PC2 los mayores pesos son educación (0,57), el ingreso total familiar (0,53) y horas trabajadas (0,43) con peso positivo, frente a edad (-0,26) y edad² (-0,35) con peso negativo (miembros del hogar tiene un peso menor -0,15). Por lo tanto, PC1 explica la mayor parte de la variabilidad y está fuertemente ligado a la edad y el tamaño del hogar, mientras PC2 agrega una dimensión de educación e ingreso y horas trabajadas.

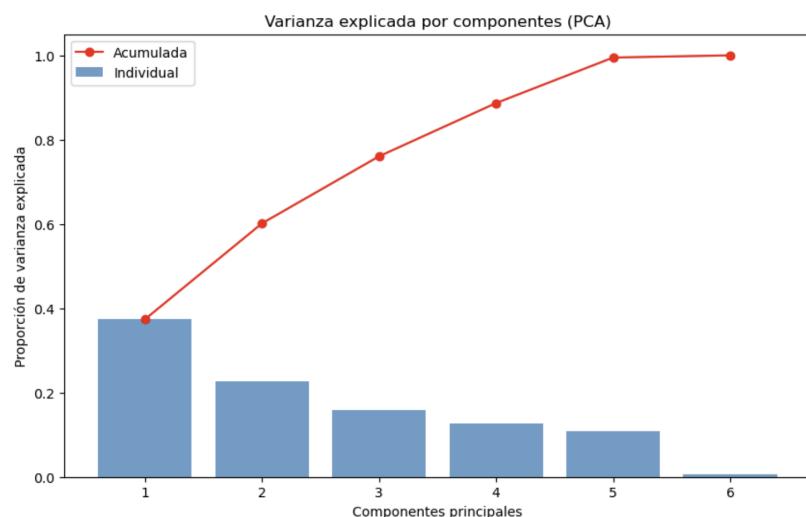
Figura 6



4. En esta consigna, realizamos un gráfico que muestra la proporción de la varianza explicada para cada uno de los seis componentes, y además, también incluimos la varianza acumulada para complementar la interpretación (Figura 7).

Las barras azules indican la proporción de varianza explicada por cada componente, la línea roja indica la varianza acumulada. El primer componente principal (PC1) explica el 37,3% de la varianza total, el PC2 explica un 22,7%, juntos indican un 60% de la varianza total acumulada. El tercer componente aporta un 15,9%, el cuarto un 12,6%, y el quinto un 10,8%, lo cual acumula casi el 100% (99,4%). El sexto componente aporta una mínima varianza del 0,52%. Con estos resultados, podemos decir que ya los primeros dos o 3 componentes capturan gran parte de la varianza de los datos, e incluir los primeros 5 garantiza aproximadamente toda la varianza.

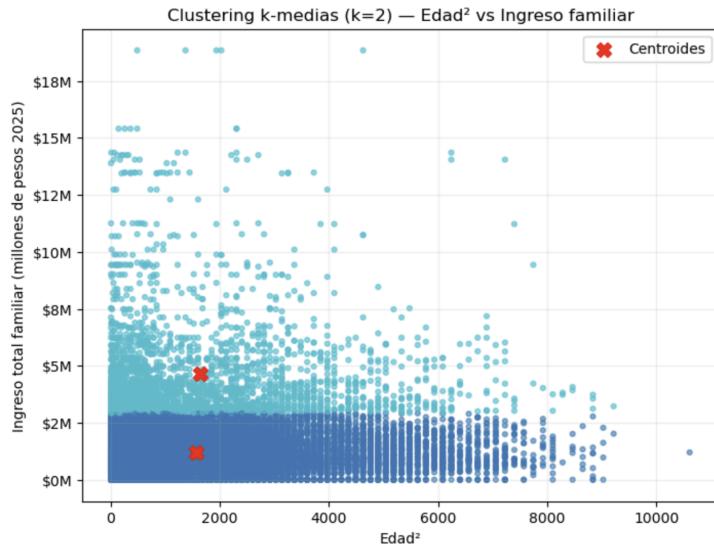
Figura 7



5.a) En esta consigna, se aplicaron k-means con 3 valores de k (2, 4 y 10) y graficamos los resultados usando edad e ingreso familiar (en millones de pesos). En los análisis de clustering se utilizó edad^2 y se normalizaron las variables para garantizar comparabilidad.

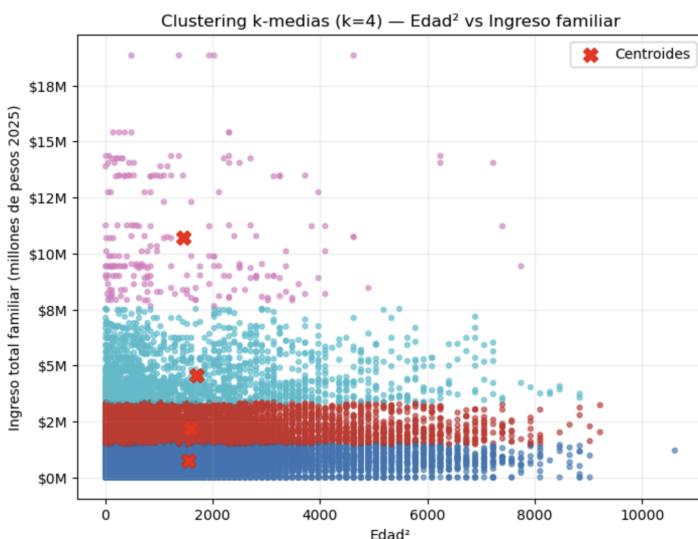
Interpretación con k=2 (Figura 8): Con k=2 (Figura 8), el algoritmo produce dos grupos principalmente por nivel de ingreso, con edades promedio similares. Esto correlaciona con pobreza/no pobreza, pero no la reproduce: no incorpora tamaño del hogar ni la CBT regional; por lo tanto no separa “correctamente” pobres de no pobres. Al comparar contra el dummy de pobreza, esperamos un accuracy moderado (no perfecto) y errores alrededor del umbral.

Figura 8



Interpretación con k = 4 clusters: En el modelo de k -medias con $k=4$ (Figura 9), el algoritmo agrupa principalmente según el ingreso total familiar, mientras que la edad² tiene poca influencia. Se distinguen cuatro niveles de ingreso: muy bajos, bajos, medios y altos, donde los centroides se alinean casi verticalmente, lo que indica que las diferencias entre grupos se explican sobre todo por el ingreso y no por la edad. En este sentido, el modelo refleja distintos estratos socioeconómicos, pero no separa con precisión a las personas pobres y no pobres, ya que no considera factores como el tamaño del hogar o la canasta básica regional.

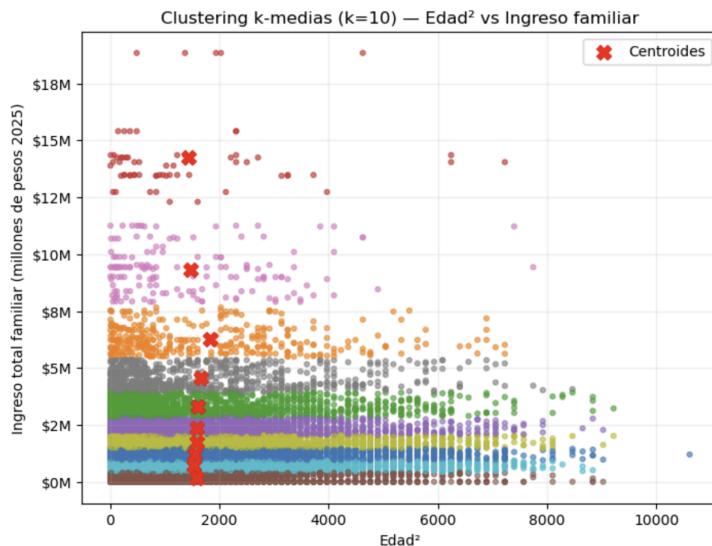
Figura 9



Interpretación con k=10: En el modelo de k -medias con $k=10$, el algoritmo genera una segmentación mucho más detallada, donde los grupos se diferencian casi exclusivamente por el nivel de ingreso total familiar, mientras que la edad² apenas influye. Los centroides se

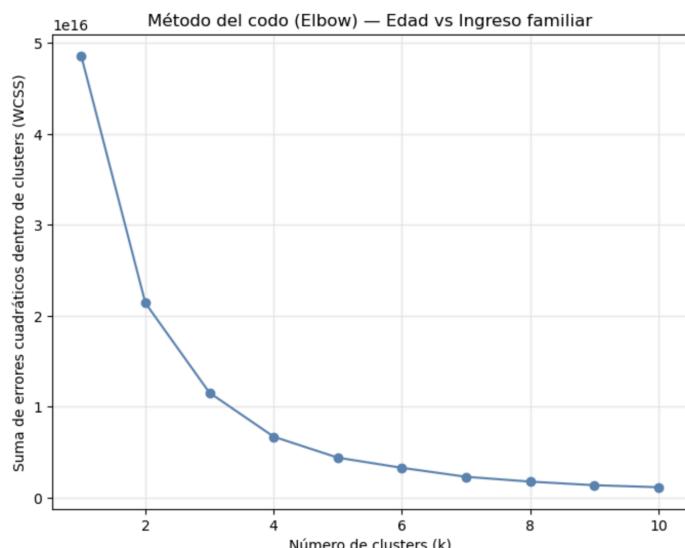
distribuyen de forma vertical, reflejando cortes progresivos en los distintos tramos de ingreso (de muy bajos a muy altos). Este resultado muestra una mayor granularidad en los estratos socioeconómicos, pero no una separación etaria clara ni una distinción precisa entre pobres y no pobres, ya que el criterio de agrupamiento sigue siendo puramente estadístico y no incorpora la línea de pobreza oficial.

Figura 10



5.b) Interpretación de la Figura 11: se observa que la reducción de la suma de errores cuadráticos es muy fuerte al pasar de $k=1$ a $k=2$, y vuelve a caer de forma marcada hasta $k=3$ y $k=4$. A partir de $k=5$ la pendiente se aplana, lo que indica que el beneficio de aumentar el número de clusters se vuelve marginal. El “codo” se ubica en torno a $k=3$ o $k=4$, que serían valores razonables para obtener una partición equilibrada. Con $k=2$ los clusters logran una separación inicial, pero aún limitada, mientras que con $k=3$ o $k=4$ se capta mejor la heterogeneidad de la población y se distinguen mejor distintos perfiles socioeconómicos.

Figura 11

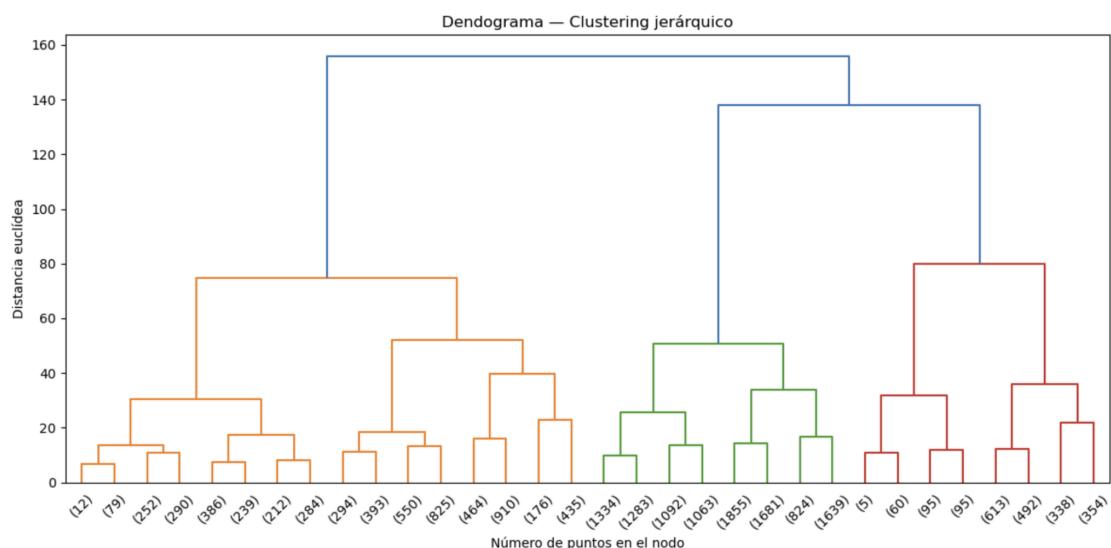


6. Un dendrograma es un diagrama que representa cómo las observaciones se agrupan progresivamente según su similitud, usando un criterio de distancia (en este caso, euclídea). Permite ver los subgrupos obtenidos para cada número posible de clusters.

Interpretación de la Figura 12: el eje vertical muestra la distancia euclídea a la que se unen los grupos y cuanto más alto se genera la unión, más diferentes son los hogares entre sí. Las ramas más cercanas a la base muestran hogares más similares entre sí, con poca variabilidad interna, mientras que las uniones más altas representan conglomerados más heterogéneos. El dendograma refleja que la población del NOA en la EPH es heterogénea, ya que existen diferencias claras en los perfiles socioeconómicos. Visualmente se distinguen tres grandes conglomerados: uno que reúne hogares con características más homogéneas, otro que combina grupos intermedios y un tercero que agrupa hogares con mayores diferencias en los valores de las variables.

Utilizamos la función “`truncate_mode=level`”, para agrupar las ramas inferiores mostrando nodos agregados con la cantidad de observaciones que representa cada uno, ya que con una gran cantidad de observaciones, los IDs en el eje horizontal se superponen y dificultan la lectura, generando “ruido” visual. Además, ajustamos el tamaño de las etiquetas para que sean más legibles. De este modo, se puede comprender la estructura general de los cluster y facilitar la interpretación visual.

Figura 12



7. Con $k=2$, el cluster 0 tiene 49,4% de no pobres y 50,6% de pobres y el cluster 1 tiene 54% de no pobres y 46% de pobres, esto significa que con 2 clusters el algoritmo no logra separar claramente entre pobres y no pobres por que ambos grupos quedan bastantes mezclados en cada división (casi un 50/50). A medida que aumenta “ k ”, aparecen clusters con mayor proporción de pobres o de no pobres, lo que significa que el algoritmo detecta distintas combinaciones de características socioeconómicas más finas que pueden asociarse a la pobreza, pero no de forma exacta.

Los resultados de las tablas 3, 4 y 5 (Ver Anexo 2), indican los resultados con la distinta cantidad de clusters. El algoritmo K-Modes asignó cada observación a un cluster en función de la similitud de sus características. Después, comparamos cada cluster con la condición “pobre=1” y “no_pobre=0”, y los porcentajes muestran la proporción de pobres y no pobres dentro de cada cluster.

Apéndice

Link al GitHub: <https://github.com/isanicola/CC408-Grupo-T3-8/tree/main/TP2>

Diccionario para la base de datos 2005

https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_disenoreg_T1_2005.pdf

Diccionario para la base de datos 2025

https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_1T2025.pdf

Anexo 1: Tabla de estadísticas descriptivas para la variable “horastrab”

Tabla 2

Estadísticas de horas trabajadas

Promedio	29.9
Desvío Estándar	24.2
Mínimo	0 horas
Mediana (p50)	30 horas
Máximo	126 horas
Cantidad de observaciones	4.236

Anexo 2:

Tabla 3

Cluster	No pobres (0)	Pobres (1)
0	49,4	50,6
1	54	46

Nota: Resultados de Cluster K-moda con K=2 teniendo en cuenta el porcentaje de pobres y no pobres según cada cluster.

Cluster	No pobres (0)	Pobres (1)
0	62,1	37,9
1	61,7	38,3
2	37,4	62,6

Tabla 4

Nota: Resultados de Cluster K-moda con K=4 teniendo en cuenta el porcentaje de pobres y no pobres según cada cluster.

Tabla 5

Cluster	No pobres (0)	Pobres (1)
0	44,6	55,4
1	49,1	50,9
2	63,7	36,3
3	48,5	51,5
4	55,9	44,1
5	49,7	50,3
6	49,4	50,6
7	64,2	35,8
8	40,1	59,9
9	29,7	70,3

Nota: Resultados de Cluster K-moda con K=10 teniendo en cuenta el porcentaje de pobres y no pobres según cada cluster.