



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



Maestría en Economía Aplicada

Materia:
Taller de Programación

Docente:
Noelia Romero

Informe – Trabajo Práctico # 2

**Histogramas, Kernels & Métodos No
Supervisados usando la EPH**

Grupo # 5:

**Cammisi, Andrés
Porco, Matías
Pineda, David**

Link GitHub: https://github.com/DavidPT0902/Grupo_5_UBA_2025.git

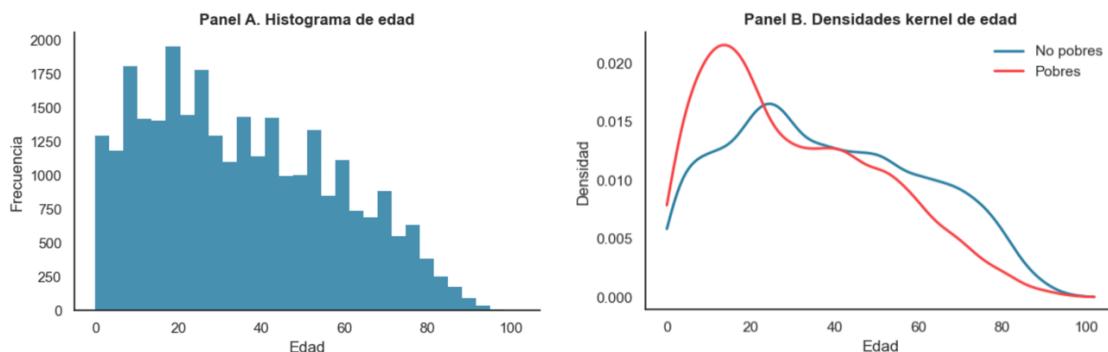
Octubre, 2025

INTRODUCCIÓN

Este trabajo analiza la evolución reciente de la pobreza y su entramado demográfico-educativo en la región Pampeana combinando estadística descriptiva y métodos no supervisados. Se construye una base armonizada para 2005T1 y 2025T1 a partir de la EPH, con ingresos expresados en pesos de 2025 y una definición de pobreza consistente basada en CBT por adulto equivalente (AE) y tamaño de hogar equivalente. El objetivo es doble: (i) caracterizar los cambios en edad, educación, ingresos y horas trabajadas, y (ii) explorar factores latentes y segmentos socioeconómicos que trasciendan la dicotomía pobre/no pobre.

PARTE I. CREACIÓN DE VARIABLES, HISTOGRAMAS, KERNELS Y RESUMEN DE LA BASE DE DATOS FINAL

A continuación se presentan los resultados más relevantes del análisis exploratorio realizada ala base consolidada de la región Pampeana para los años 2005 y 2025, la cual se elaboró a partir de los datos de la EPH.



El panel A muestra que la distribución de la población por edades es asimétrica a derecha (es decir, existe una mayor concentración de personas en los tramos de edad más jóvenes). Como muestra el panel B, este sesgo es mayor en el caso de las personas pobres (lo que refleja una alta incidencia de la pobreza en niños y jóvenes); y menor en el caso de las personas no pobres (cuya distribución muestra un mayor peso relativo de los adultos y adultos mayores).

Tabla 1. Región Pampeana: Cantidad de años de educación (2005 y 2025)

Año	Mean	Std	Min	P50	Max
2005	8.72	4.7	0.0	8.0	20.0
2025	9.98	4.67	0.0	12.0	20.0

Como se indica en la Tabla 1, para el 2005 la región Pampeana presentó un promedio de 8.72 años de educación con un incremento a 9.98 años en 2025, lo que reflejaría mejoras educativas en esta región. De manera aún más significativa, la mediana pasó de 8 años de educación en 2005 a 12 años en 2025 (producto, en buena

medida, de la obligatoriedad de la educación secundaria instaurada en el año 2006). En ambos años la desviación estándar es significativa (del orden de 4,7 años), lo que indica que coexisten personas con reducida o nula educación formal y otras con formación a nivel superior.

En cuanto a la distribución de los ingresos familiares, el panel A muestra también una distribución asimétrica a derecha (es decir, existe una mayor concentración de personas en los estratos de menores ingresos). Como muestra el panel B, este sesgo es mayor en el caso de las personas pobres (que, por definición, tienden a concentrarse por debajo del umbral de pobreza) y menor en el caso de las personas no pobres (donde ocurre lo contrario). Dado que estamos analizando la distribución de *personas* por niveles de *ingreso total familiar*, lo que llamamos aquí “línea de pobreza” es, en rigor, la mediana del ingreso total familiar necesario para no ser pobre.

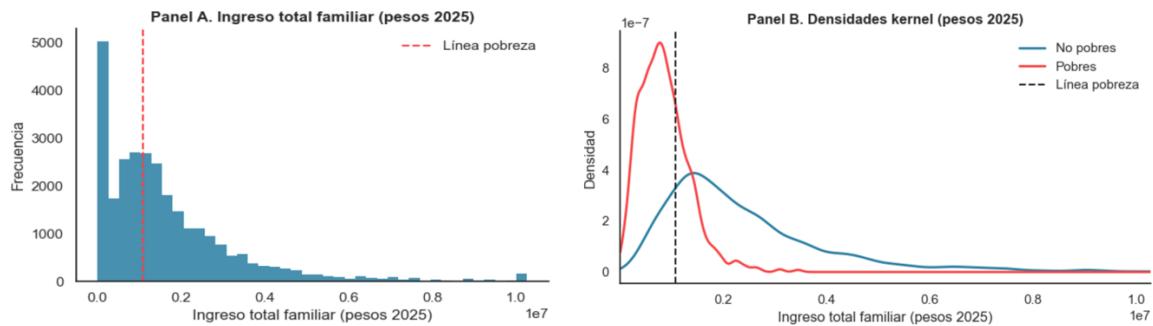


Tabla 1. Región Pampeana: estadística descriptiva de horas trabajadas

Año	Mean	Std	Min	P50	Max
2005	41.59	22.35	0.0	44.0	112.0
2025	35.43	19.07	0.0	40.0	112.0

Fuente: elaboración propia a partir de EPH (2005) & EPH (2025).

En el área laboral se registra que para el 2005 la horas trabajadas promedio de los jefes de hogar era aproximadamente 41 horas y para el 2025 de 35 horas. Estas cifras pueden estar asociadas con una reducción del empleo formal y, a su vez, un incremento de la inactividad o subocupación laboral.

Tabla 2. Región Pampeana: Resumen de la base unificada

Año	Cantidad de observaciones	Observaciones NA en “pobre”	Cantidad de pobres	Cantidad de “no pobres”	Variables tratadas
2005	14,651	0	4,364	10,287	42
2025	13,803	0	7,733	6,070	42
Total	28,454	0	12,097	16,357	42

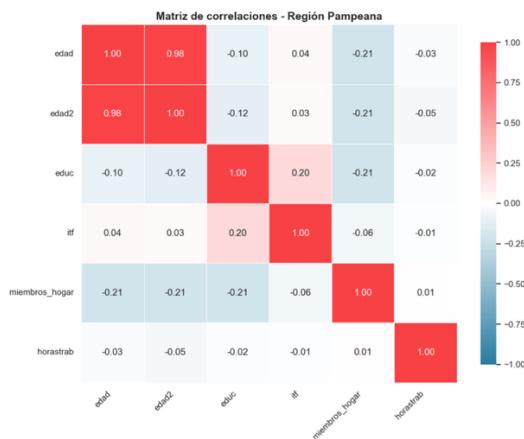
Fuente: elaboración propia a partir de EPH (2005) & EPH (2025).

Finalmente, se destaca que la base final de la región Pampeana contiene 28,454 observaciones: 14,651 para 2005 y 13,803 para 2025. La variable “pobre” no presentó

observaciones faltantes. Los datos muestran un incremento importante de personas pobres, de 4,364 a 7,733 en los dos años de análisis, lo que indica un deterioro en las condiciones de vida de la población, tomando como referencia el ingreso familiar equivalente. En total se trabajaron 43 variables en cada año, las cuales fueron recodificadas hasta alcanzar su homogenización. Estas regularidades (hogares más jóvenes y numerosos entre los pobres, mayores años de educación y mejores ingresos en el polo opuesto) sugieren la existencia de dos ejes importantes: uno demográfico y otro socio-educativo-ingresos. A continuación formalizamos esta intuición con correlaciones y PCA.

PARTE II. MÉTODOS NO SUPERVISADOS

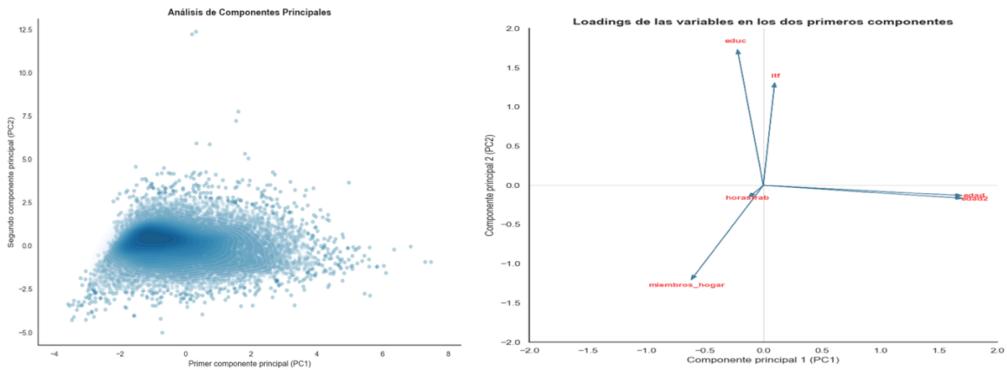
Para profundizar en el análisis se realizaron ejercicios de PCA y clusters. El detalle de los principales hallazgos se presenta en los siguientes párrafos.



Las correlaciones son, en general, bajas o moderadas. Tal como cabría esperar, se observa una relación positiva entre horas trabajadas e ingreso familiar, y entre educación e ingreso familiar. Por otra parte, se observa una relación negativa entre la cantidad de miembros del hogar y el ingreso total familiar (lo que indica que los hogares pobres tienden a ser más numerosos); y también una relación negativa entre la cantidad de miembros del hogar y la edad (lo que sugiere que los hogares numerosos se caracterizan por una mayor presencia de niños y personas jóvenes).

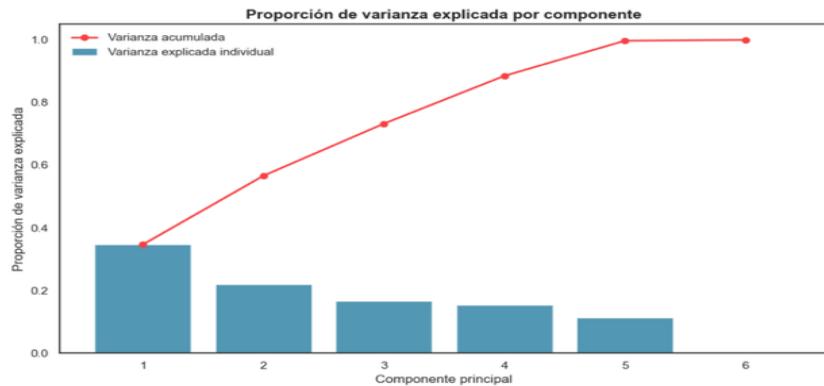
PCA

Utilizamos PCA para condensar la información de las seis variables en **factores latentes** que expliquen la mayor parte de la variación conjunta sin imponer supuestos causales. El gráfico de dispersión de PCA muestra que los dos primeros componentes explican el 56.6% de la varianza total. El primer componente explicar el 34.6% de la varianza y el segundo componente el 22.0%, en total. La concentración sugiere una estructura relativamente homogénea, con pocos casos atípicos.



Analizando los loadings, podemos observar que:

- **PC1 (demográfico):** edad y edad2 con cargas **positivas**; miembros_hogar con carga **negativa**. Este componente captura un **gradiente de ciclo de vida / tamaño de hogar**.
- **PC2 (socio-educativo-ingresos):** educ e itf con cargas **positivas** y, nuevamente, miembros_hogar con **negativa**. Este eje resume **acumulación educativa y posición de ingresos**.



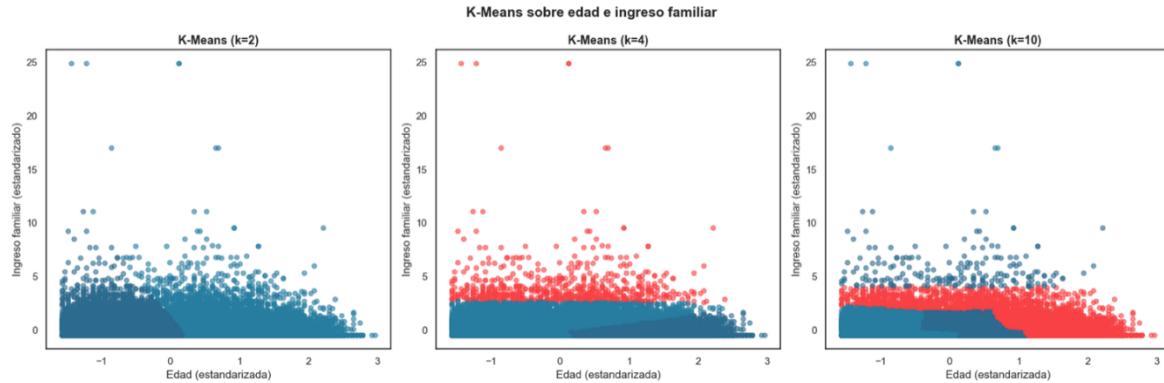
En resumen, en el primer componente la variable de mayor peso en la edad y en el segundo componente parecen tener mayor influencia las variables nivel educativo e ingreso total familiar. Si la heterogeneidad se organiza en (al menos) dos ejes, una partición “parsimoniosa” debería tener **pocos clústeres**. Evaluamos K-means y jerárquico para ver si dicha segmentación resulta estable e interpretable.

K-means

Dibujamos un mapa con edad e ingreso familiar (itf) estandarizados, así ninguna variable “pisa” a la otra, y dejamos que K-means encuentre perfiles típicos: por ejemplo, jóvenes de ingreso bajo, edad media con ingreso medio, mayores con ingreso alto, etc. Después evaluamos si esa división ayuda a entender el ciclo de vida y

distinguir niveles socioeconómicos (y elegimos cuántos grupos conviene con Elbow y Silhouette).

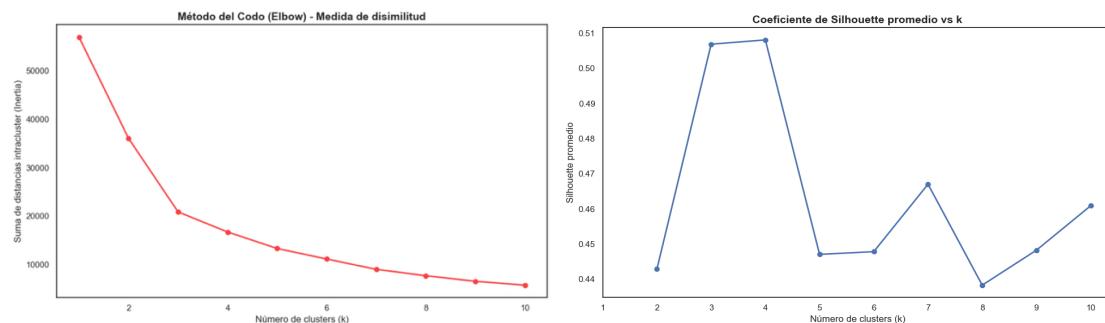
Con **K-means** sobre (edad, ift) estandarizadas, **k=2** separa sobre todo por **ingresos**; la alineación con pobre/no pobre es buena cerca del umbral, pero no perfecta (zona gris esperable). Con **k=4** aparecen **estratos** más nítidos que combinan ingreso y ciclo de vida; con **k=10** la granularidad aumenta a costo de interpretabilidad.



Los criterios de validación sugieren un **punto óptimo alrededor de k=4**:

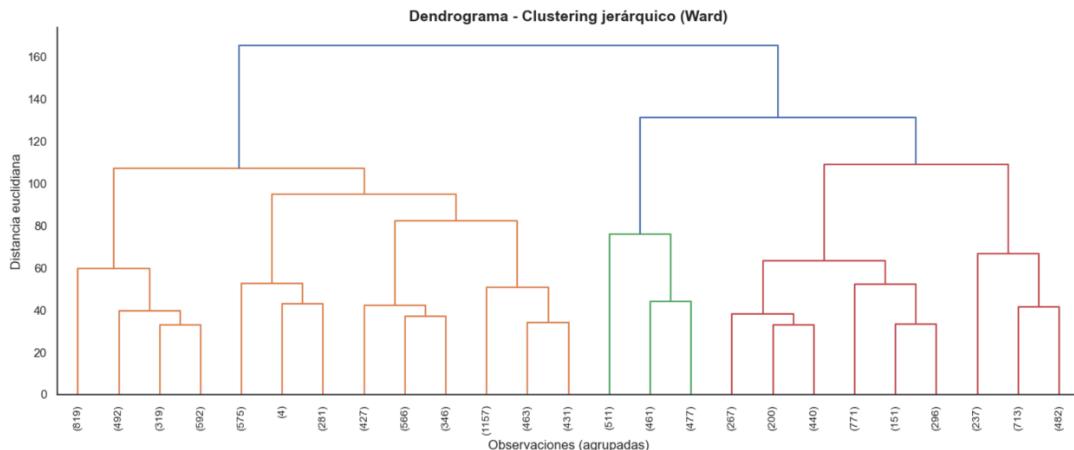
- **Elbow** muestra **codo en $k \approx 3-4$** (inercia vs k)
- **Silhouette** alcanza **máximo** en torno a **k=4**

La convergencia de ambos criterios refuerza que **cuatro grupos** capturan la heterogeneidad con buen equilibrio entre parsimonia e interpretación.



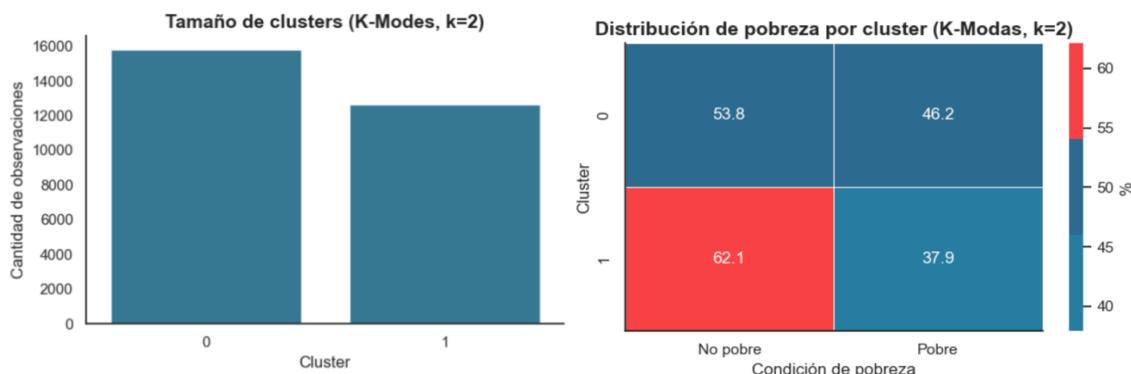
Cluster Jerárquico

Incorporamos clúster jerárquico ya que posee la ventaja de no fijar k de antemano: muestra la secuencia de fusiones y los “saltos” de heterogeneidad entre grupos. El dendograma obtenido mediante el método de Ward distingue tres conglomerados principales, unidos a diferentes niveles de distancia, indicando cierta variabilidad entre ellos. Por su parte, en la parte baja se representan observaciones similares entre sí.



K modas

K-modas es la versión para variables categóricas de K-means: en lugar de promedios usa las modas (categorías más frecuentes) para construir los “prototipos” de cada grupo y mide la disimilitud por coincidencias/desaciertos de categorías. Al estimar un modelo k-modas se lograron identificar dos clusters. El cluster 0 agrupa casi 15,791 observaciones y el cluster 1 agrupa 12,663. En el cluster 0, la proporción de personas pobres y no pobres es equilibrada, sin embargo, se observa un leve predominio de los no pobres (53.8%). En el cluster 1, las diferencias son más claras: el 62.1% son no pobres. La distribución relativamente balanceada entre grupos indica que el modelo logra una segmentación coherente y estable, capturando heterogeneidad en la población sin imponer divisiones artificiales, lo que permite detectar perfiles socioeconómicos, aunque con reducido poder discriminante en relación con la condición de pobreza



Conclusiones

El análisis muestra una pérdida de bienestar entre 2005T1 y 2025T1, con aumento notorio de la pobreza a pesar de avances educativos. La heterogeneidad socioeconómica se organiza principalmente en dos factores—demográfico y socio-educativo-ingresos—que, al combinarse, justifican una segmentación en cuatro clústeres interpretables. Esta lectura sugiere que las políticas deberían ir más allá del umbral de pobreza y focalizar por perfiles, atendiendo simultáneamente composición del hogar, ciclo de vida, acumulación educativa e inserción laboral.