



UBA FCE
Universidad de Buenos Aires
Facultad de Ciencias Económicas

Big Data y Machine Learning (UBA) - 2025

Trabajo práctico 3: Histogramas, Kernels & Métodos No Supervisados usando la EPH

Grupo 1

Francisco Ariel Gonzalez Häberle 903012

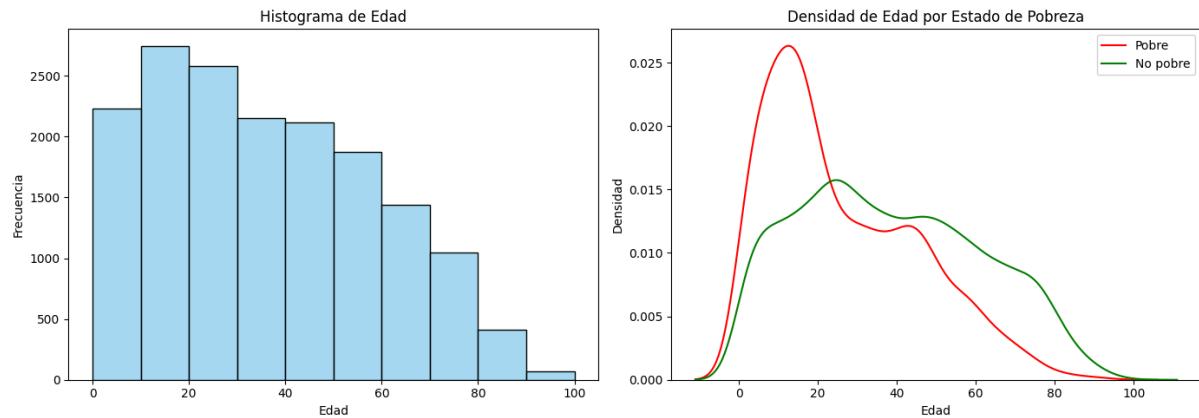
David Jimenez Jaldin 909323

Lourdes Sofia Beltramo 888198

Link github: <https://github.com/davidjij14082000-png/BigDataUBA-Grupo-1>

Parte I: Creación de variables, histogramas, kernels y resumen

Ejercicio 1



El panel A muestra el histograma de la variable edad, donde se observa una distribución concentrada en edades jóvenes y adultas, con mayor reiteración en el rango de 10 a 30 años y una disminución progresiva hacia las edades más avanzadas. La distribución tiene asimetría hacia la derecha, común en poblaciones con mayor proporción de personas jóvenes.

En el panel B, que presenta las densidades de edad por estado de pobreza, se aprecia que la población pobre (línea roja) está más concentrada en los grupos de menor edad, mientras que la población no pobre (línea verde) se distribuye de manera más amplia y con mayor presencia en edades adultas y avanzadas. En conjunto, los gráficos sugieren que la pobreza afecta en mayor medida a los grupos más jóvenes, mientras que los no pobres predominan en edades medias y mayores.

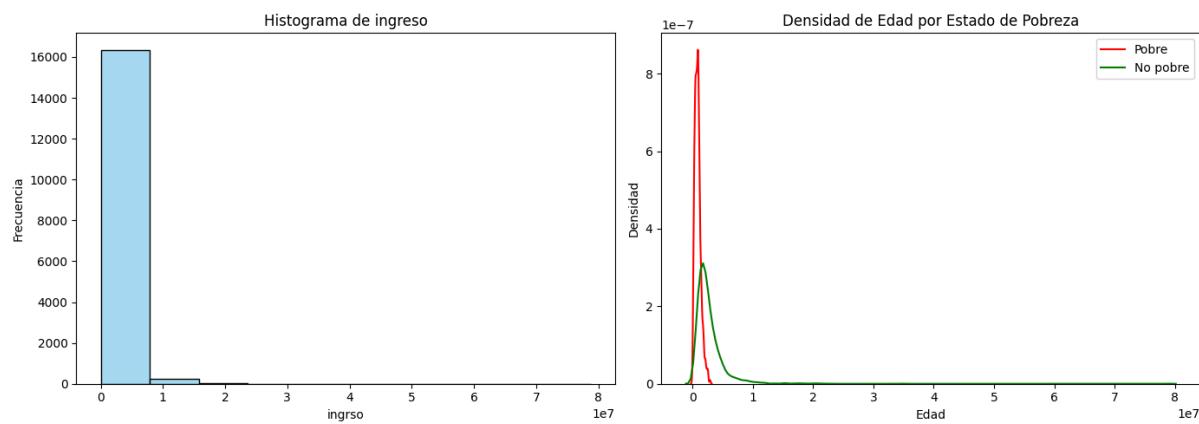
Ejercicio 2

Estadística descriptiva sobre la variable *educ*: La variable *educ* presenta una tendencia central en torno a la educación secundaria completa, con un alto nivel de dispersión y la existencia de personas que no han cursado educación formal así como otras que tienen estudios universitarios o de posgrado.

	educ
count	15,692.00
mean	12.28
std	5.24
min	0.00

25%	8.00
50%	14.00
75%	17.00
max	24.00

Ejercicio 3



El panel A muestra el histograma del ingreso total familiar (ITF ajustado a pesos de 2025), donde se observa una fuerte concentración de hogares con ingresos bajos y la existencia de pocos hogares con ingresos muy elevados. La distribución es muy asimétrica, reflejando la desigualdad en la distribución del ingreso.

En el panel B, las curvas de densidad por condición de pobreza dan cuenta que los hogares pobres (línea roja) se agrupan en los tramos de ingreso más bajos, mientras que los no pobres (línea verde) se desplazan hacia niveles de ingreso mayores. En conclusión, ambas representaciones confirman una brecha significativa entre pobres y no pobres.

Ejercicio 4

La tabla muestra las estadísticas descriptivas de la variable horas trabajadas. En promedio, las personas trabajan alrededor de 26,4 horas semanales, aunque la alta desviación estándar (83,6) indica una gran dispersión en los datos. El valor mediano es de solo 5 horas, lo que sugiere que muchas personas trabajan pocas horas o ninguna (posiblemente desempleados). Sin embargo, el valor máximo de 1.998 horas refleja la presencia de valores atípicos, probablemente errores de registro o casos excepcionales.

	horas_trabajadas
count	12,882.00
mean	26.43
std	83.56
min	0.00
25%	0.00
50%	5.00
75%	40.00
max	1,998.00

Ejercicio 5

	2005	2025	Total
Cantidad observaciones	9484	7181	16665
Cantidad de observaciones con NAs en la variable “Pobre”	113	2872	2985
Cantidad de Pobres	2532	1341	3873
Cantidad de No Pobres	6839	2968	9807
Cantidad de Variables	33	33	33

La tabla muestra un resumen general de la calidad de los datos y las observaciones en las bases que corresponden a los años 2005 y 2025. En total, hay 16.665 observaciones, repartidas entre 9.484 en el año 2005 y 7.181 en el año 2025.

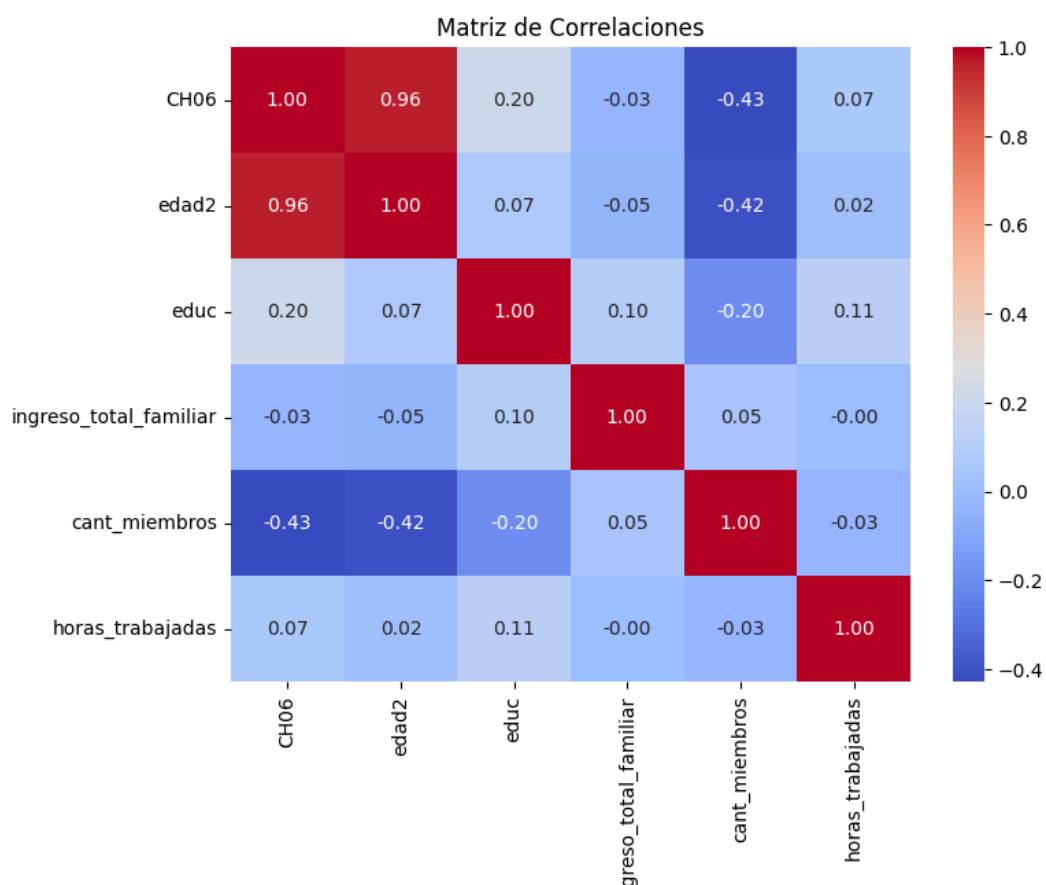
En la variable *Pobre*, se ve un incremento en los valores faltantes (NAs) en 2025 (2.872 casos) con respecto a 2005 (113 casos), lo que podría tener un impacto en el análisis de pobreza para ese año.

Respecto a la distribución, se observan 3.873 personas pobres y 9.807 no pobres, lo que indica que en la muestra total hay una proporción más elevada de no pobres.

Parte 2: Métodos no supervisados

Ejercicio 1

La educación tiene una correlación positiva leve con la edad y los ingresos familiares, lo que sugiere que a medida que se sube el nivel educativo, también aumentan los ingresos. La cantidad de miembros en el hogar tiene una correlación negativa con la educación y la edad, lo que indica que las familias con mayor número de integrantes tienden a tener jefes de hogar más jóvenes o menos educados. Por último, las horas trabajadas no tienen una correlación significativa con ninguna variable, lo que sugiere que la relación con los ingresos, la edad o la educación es débil.

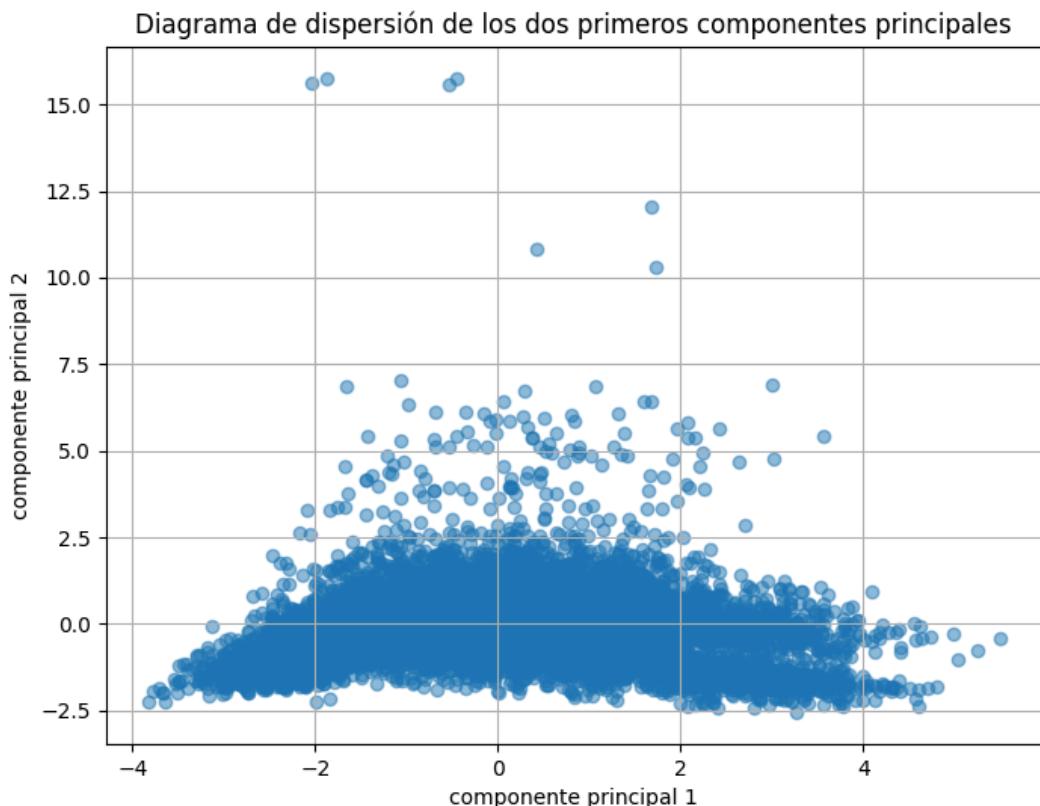


PCA

ejercicio 2

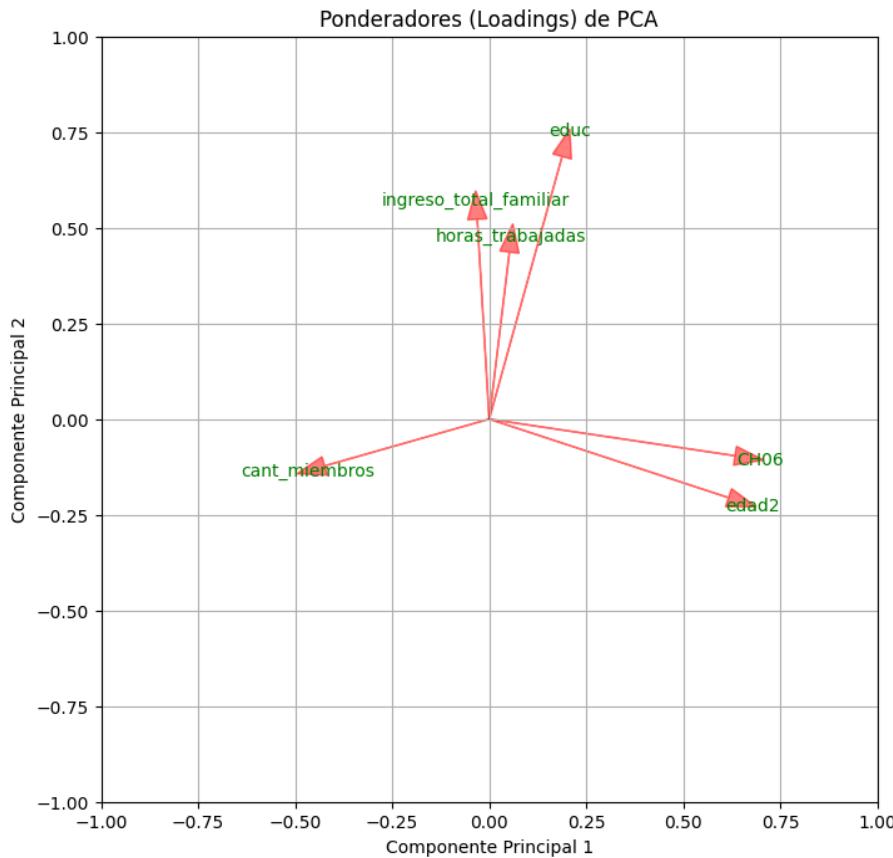
El diagrama de dispersión del PCA muestra cómo se distribuyen las personas según los dos componentes principales. La mayoría de los puntos están concentrados cerca del centro, lo que indica que la mayoría de los hogares tienen características similares. El primer componente explica la mayor parte de las diferencias entre las personas, mientras que el segundo aporta menos.

variación. Se observan algunos casos atípicos en la parte superior del gráfico, que representan hogares con ingresos, educación o trabajo mucho más altos que el promedio.



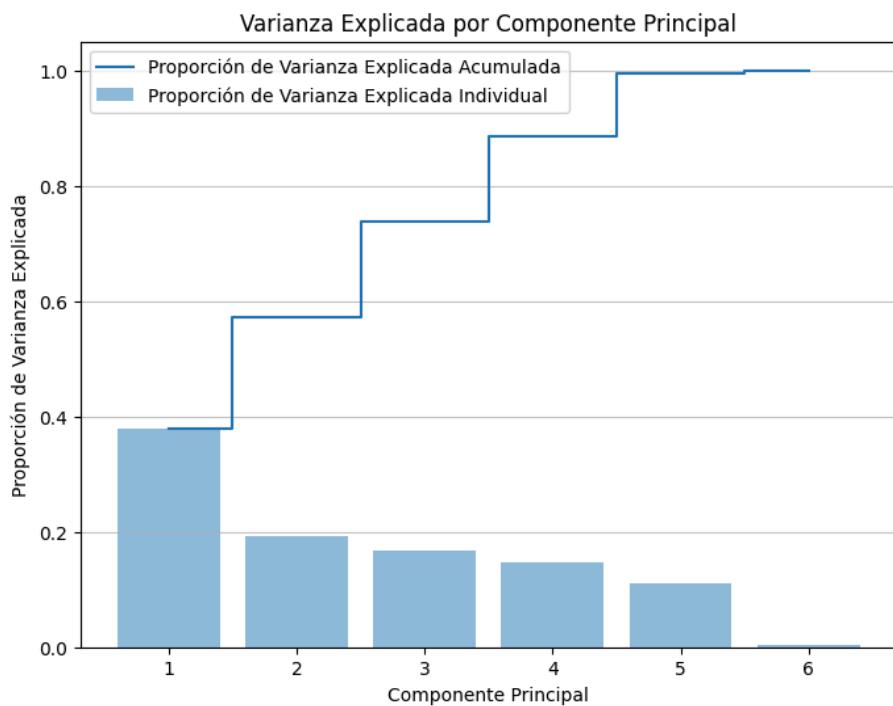
Ejercicio 3

El gráfico de componentes principales muestra dos dimensiones principales. El primer componente diferencia a los individuos según su edad y tamaño del hogar, donde las personas de mayor edad suelen vivir en hogares más pequeños. El segundo componente está asociado con el nivel educativo, los ingresos y las horas trabajadas, reflejando el nivel socioeconómico y laboral, lo cual sugiere que más educación suele llevar a más horas trabajadas y a mayores ingresos.



Ejercicio 4

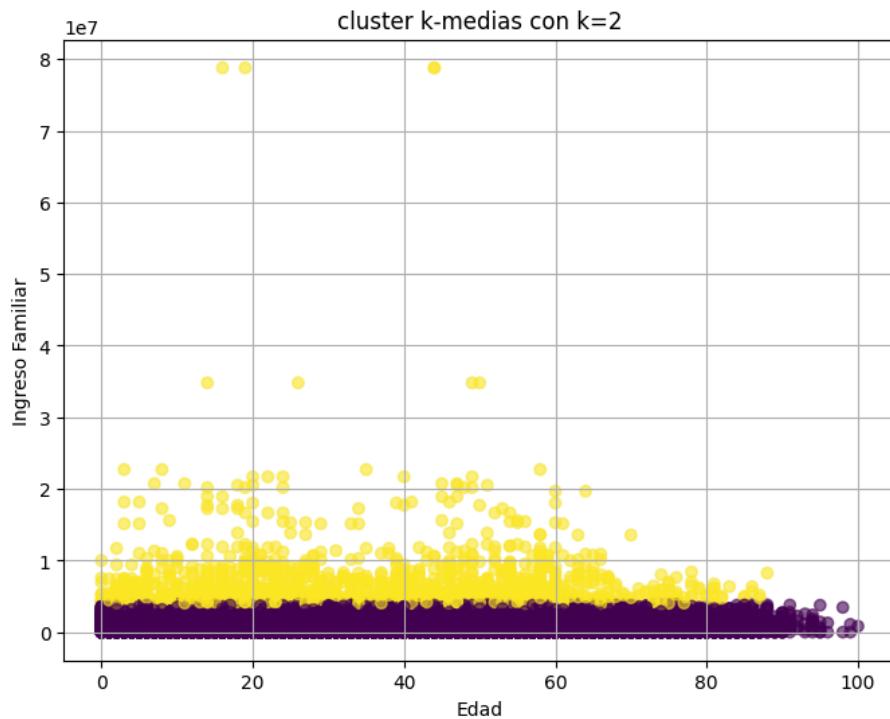
El componente principal 1 es por lejos el que más explica, cerca del 38% de la varianza total de los datos por sí mismo. el segundo cerca del 19% y el tercero alrededor del 17%. Focalizando en la varianza acumulada podemos observar que los dos primeros componentes explican el 57% de la varianza total. Para alcanzar el 90% se necesitan al menos cuatro componentes.



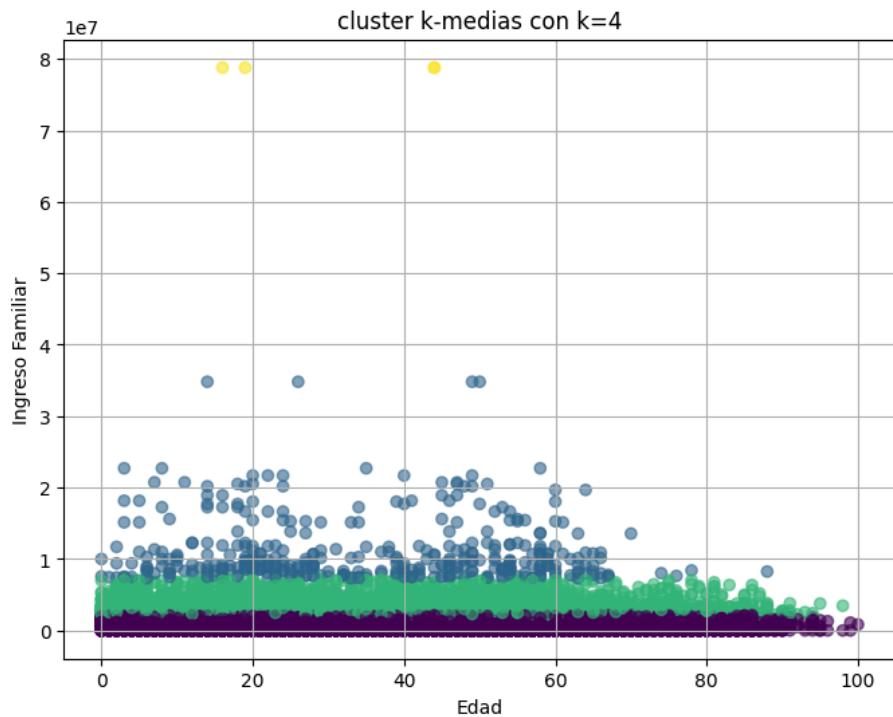
Cluster

Ejercicio 5

El algoritmo con $k=2$ logra separar de manera efectiva a las personas "pobres" y "no pobres" si la pobreza se define mediante un umbral de ingreso fijo muy bajo. El clúster violeta representa la población que cae por debajo de ese umbral, mientras que el clúster amarillo representa al resto.



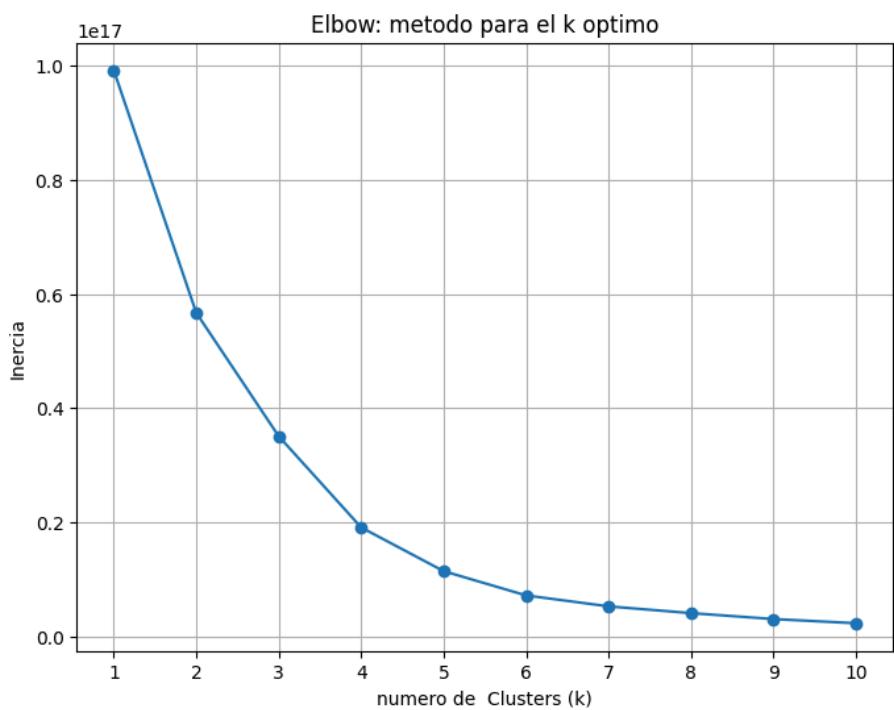
Al aumentar a $k=4$, el modelo comienza a realizar más segmentaciones dentro de los datos de bajos ingresos. El clúster de ingresos más bajos violeta sigue muy concentrado en el eje inferior, pero aparecen nuevos clústeres (verde claro, azul) que segmentan a la población de ingresos bajos a medios y altos.



Con $k=10$, la segmentación en la región de ingresos bajos y medios se vuelve mucho más fragmentada y difusa. Hay aquí una “sobre segmentación” la cual no ofrece una interpretación clara. Al haber 10 centroides hay muchos clusters con poca diferenciación.

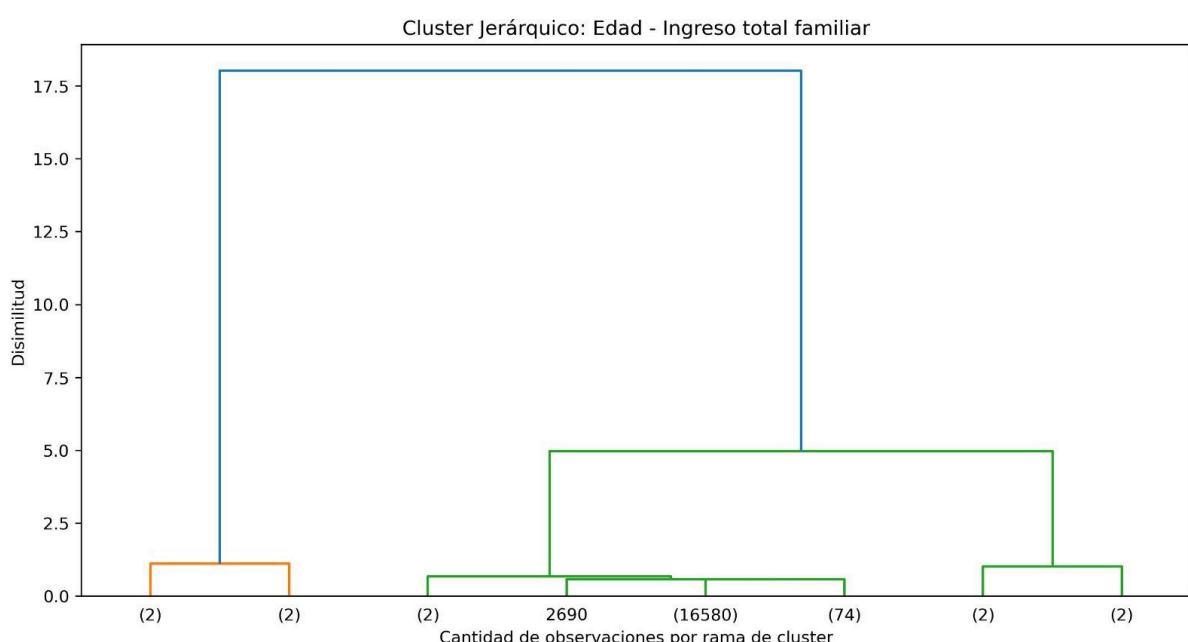


El gráfico del Método Elbow muestra cómo cambia la “inercia” al aumentar la cantidad de clusters. A medida que se agregan más grupos, la inercia disminuye, ese punto donde la curva se dobla aparece en $k = 3$, lo que indica que 3 es el número óptimo de grupos. Esto significa que dividir la población en tres grupos permitiría una mejor representación de las diferencias sin complicar demasiado el modelo(podrían ser Hogares de ingresos bajos, Hogares de ingreso medio-bajo, Hogares de ingreso medio-alto o alto)



Ejercicio 6

El dendograma o cluster jerárquico es un método no supervisado que consiste en un diagrama de árbol que se construye desde los niveles inferiores hacia los superiores. El eslabón más bajo de las ramas ("las hojas") consiste en pares de observaciones agrupadas por similitud. Para ello en este caso se calcula la distancia euclíadiana entre las observaciones (una vez estandarizadas las variables) y se agrupa las de menor distancia. Posteriormente los clusters se comienzan a agrupar entre sí hasta quedar resumidos en un único cluster. El criterio de unión utilizado en este caso para agrupar los casos es el "completo", el cual consiste en definir la distancia entre clusters a partir de la máxima distancia existente entre un par de observaciones de un cluster y otro.



Ejercicio 7

Este gráfico muestra los resultados del algoritmo K-Modas con diferentes números de clusters ($k=10$, $k=4$ y $k=2$) para agrupar los individuos según sus características socioeconómicas representadas por variables dummies.

Cuando se usan 10 clusters, se ve una distribución muy desigual, donde algunos grupos tienen más de 3000 observaciones y otros menos de 700, lo que indica una alta fragmentación. Con 4 clusters, la distribución se vuelve más equilibrada y representativa de la población.

Por último, con $k=2$, los dos grupos tienen prácticamente la misma cantidad de observaciones (8284 y 8381) es decir, el modelo logró dividir la población en dos grandes grupos de tamaño similar. Lo cual pareciera en primera instancia indicar que no permite identificar a los individuos pobres ya que la proporción de los individuos clasificados de esta manera la muestra (excluyendo aquellos para los cuales no pudimos establecer una clasificación) es del 28% (lejos de 2 grupos de igual tamaño). Sin embargo observamos que mientras que en el cluster "0" solo el 24,43% de los miembros con catalogados como pobres mientras que en el cluster "1" este porcentaje asciende al 40,12%. Si reasignamos a los individuos en cluster agrupándolos por hogar en función de cómo fueron catalogados la mayoría de los miembros de dicho hogar, vemos que la brecha se amplía (18,71% de los miembros del cluster 0 y 49,03% del 1 son pobres). A su vez observamos que el 64% de los individuos pobres fueron asignados al cluster 1, porcentaje que aumenta al 69% si consideramos la clasificación por hogar.

Finalmente si utilizamos el ser asignado al cluster 1 como un predictor de que un individuo se encuentra catalogado como pobre, obtenemos dentro de la muestra un porcentaje de predicción del 57,21%. Al realizar la agrupación por hogares la precisión es del 66,56%. Concluimos entonces que si bien el K-Modas con $k=2$ no parece separar a la población en pobres y no pobres si consideramos que existe cierto grado de relación en la forma en que el algoritmo asignó las variables categorías utilizadas y la condición de pobreza.

