**UNIVERSIDAD DE BUENOS AIRES - 2025**

# Big Data y Machine Learning

## Trabajo Práctico N° 3

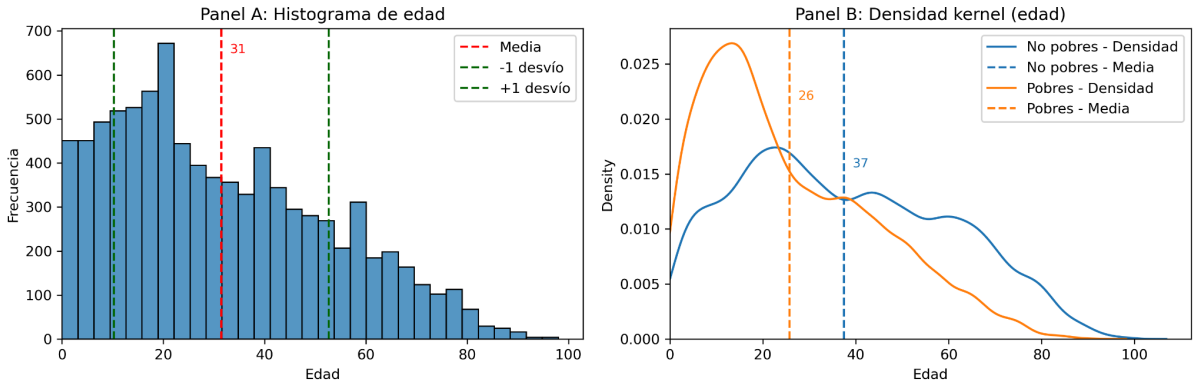
## Histogramas, Kernels & Métodos No Supervisados usando la EPH

**Fecha de entrega:** 14 de octubre a las 13:00 hs.

**Link al repositorio del grupo:**
https://github.com/StefanoPistoia/Big-Data-Grupo-2/tree/main

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

1. *Cree la variable "edad2" definida como edad$^2$ (edad al cuadrado). Presente un histograma de la variable edad en un panel A, y a la par una distribución de kernels para los pobres y no pobres en un panel B (esto es, son dos líneas de kernel en este segundo panel). Comente brevemente la distribución de edades en estos dos paneles (3-4 oraciones).*




Ambos métodos exhiben de manera relativamente similar la distribución de edad de la muestra: un grupo considerable que oscila entre los 0 y los 20 años, y luego una disminución progresiva a medida que se avanza en edad.

También pareciera que asoma una correlación inversa entre la edad poblacional promedio y el nivel socioeconómico (la edad media encuesta de los pobres es sustancialmente menor que la de los no pobres).

2. *Creación de la variable 'educ'; estadística descriptiva y comentarios de la misma.*

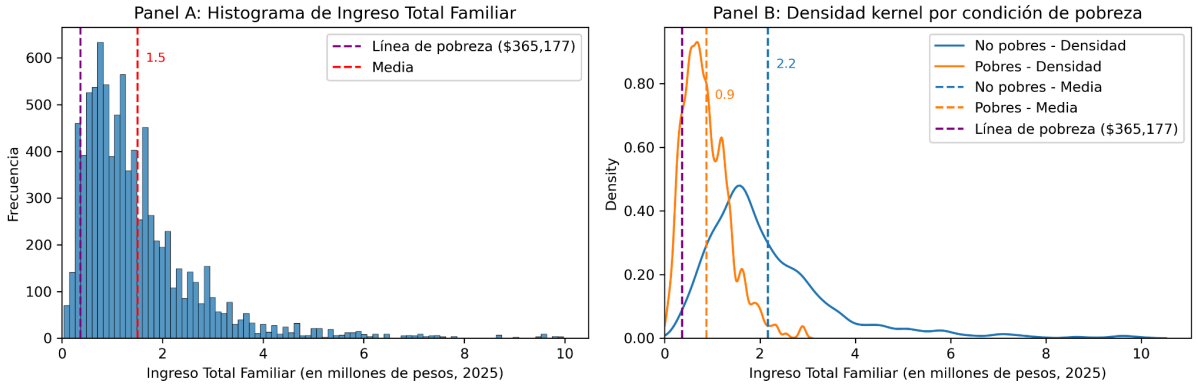
```
respondieron['educ'].describe().round()
```

```
count    8032.0
mean        9.0
std         5.0
min         0.0
25%         6.0
50%         9.0
75%        12.0
max        19.0
Name: educ, dtype: float64
```

Se evidencia que la cantidad de años promedio de educación es de 9, lo que equivale a: primaria completa (6 años) + secundaria incompleta hasta tercer año

(3 años). Asimismo, el desvío estandar es considerable, haciendo que el intervalo entre la cola inferior y superior de la magnitud del desvío sea [4,14], que en términos categóricos podría traducirse como [*cuarto grado de primaria,segundo año de terciario/universitario*].

3. *Actualización de 'ingreso_total_familiar' al poder de compra del 1T de 2025; histograma de la variable ingreso_total_familiar y las distribuciones de kernels para pobres y no pobres en un panel B; comentario sobre la distribución de ingresos en estos dos panels (3-4 oraciones).*




Se observa que un rango considerable de la muestra concentra los valores de su ingreso total familiar por debajo de la media, que es de aprox. $1.500.000.

Al desdoblarse según la condición de pobre/no pobre, para el caso de los pobres, el rango de ingresos se acota aún más. Asimismo, es para destacar que existe una parte de la población pobre cuyo ingreso total familiar se encuentra muy cerca de la moda (el pico de la función de densidad) y por debajo de la línea de pobreza individual.

4. *Para el jefe del hogar, cree la variable horastrab como el total de horas trabajadas como la suma de las horas en la ocupación principal y otras ocupaciones (PP3E_TOT + PP3F_TOT). Presente una estadística descriptiva (promedio, sd, min, p50, max) de dicha variable creada y comente.*

```
print(stats)
print()
```

```
count    2062.0
mean       31.0
std        25.0
min         0.0
25%         0.0
50%        33.0
75%        48.0
max       112.0
Name: horastrab, dtype: float64
```

Se observa una amplia magnitud del desvío respecto a la media; lo cual sugiere que en una gran cantidad de hogares el jefe de hogar se dedica a actividades del hogar y no posee un trabajo remunerado (aquellos cuyo 'min' es igual a 0).

5. *¿Cuál es el tamaño de la de la base de datos para su región con las variables originales unificadas? Para ello complete la tabla 1 que se le diseña abajo y comente.*

**Tabla 1. Resumen de la base final para la región NEA**

| | **2005** | **2025** | **Total** |
|---|---|---|---|
| Cantidad observaciones | 5406 | 3340 | 8746 |
| Cantidad de observaciones con NAs en la variable "Pobre" | 0 | 0 | 0 |
| Cantidad de Pobres | 2867 | 1637 | 3504 |
| Cantidad de No Pobres | 2539 | 1703 | 4242 |
| Cantidad de variables limpias y homogeneizadas | 34 | 34 | 34 |

Nota: Calcular la cantidad de pobres y no pobres a partir de la variable de *Pobre* que crearon en el trabajo práctico 2.

| ANO4 | Cantidad observaciones | NAs en "pobre" | Cantidad de Pobres (pobre=1) | Cantidad de No Pobres (pobre=0) | Cantidad de variables (columnas) |
|---|---|---|---|---|---|
| 2005 | 5406 | 0 | 2867 | 2539 | 34 |
| 2025 | 3340 | 0 | 1637 | 1703 | 34 |

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

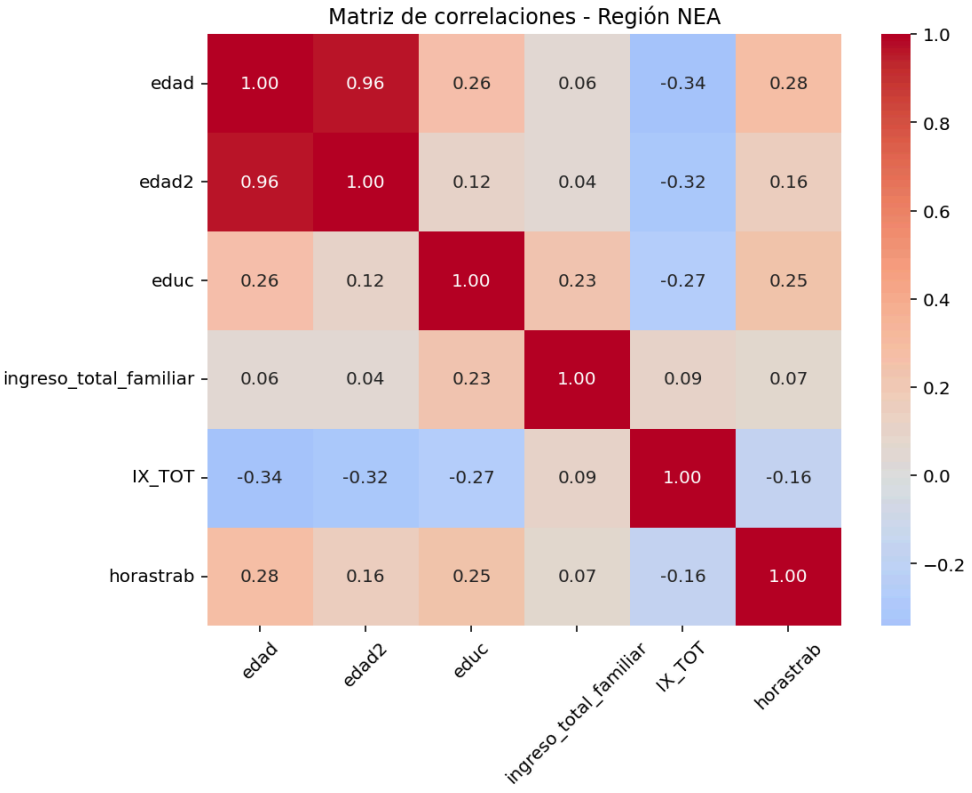
1. *Matriz de correlaciones*

Se observa una correlación muy alta entre edad y edad², esperable dado que una es el cuadrado de la otra.

La educación presenta una correlación positiva moderada con el ingreso total familiar (r = 0.23), indicando que a mayor nivel educativo, tienden a registrarse mayores ingresos.

La cantidad de miembros del hogar se asocia negativamente con casi todas las variables, en particular con la edad (r = –0.34), lo que sugiere que los hogares más grandes suelen corresponder a jefes o jefas de hogar más jóvenes.

Las horas trabajadas se relacionan débilmente de forma positiva con la edad y la educación, y no muestran una relación clara con el ingreso familiar.



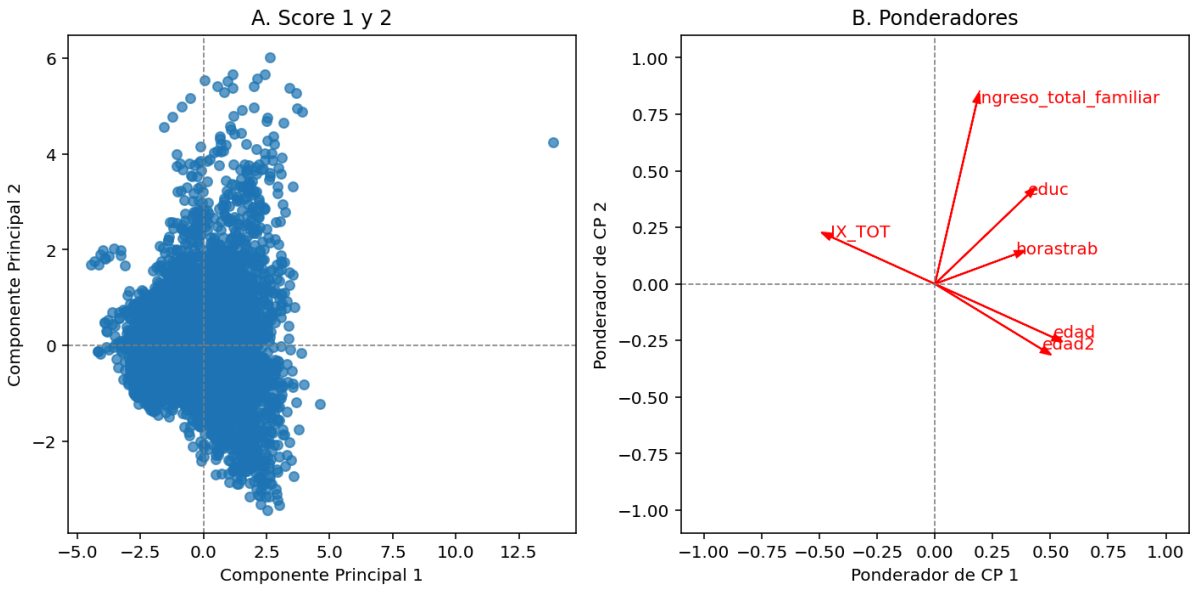

*2 y 3. PCA con ingreso; graficación de Scores y Ponderadores*

Se observa una distribución relativamente concentrada, sin grupos claramente diferenciados, lo que sugiere una estructura continua de variación más que una segmentación nítida entre los casos.

El primer componente (CP1) parece capturar principalmente las diferencias asociadas con el nivel socioeconómico, ya que, como se ve en el gráfico de ponderadores (B), está fuertemente influido por las variables ingreso total familiar, educación y, en menor medida, horas trabajadas, todas con cargas positivas.

El segundo componente (CP2), en cambio, refleja una oposición entre la cantidad de integrantes del hogar y el resto de las variables, indicando que, a igualdad de edad y educación, los hogares más numerosos tienden a ubicarse en el extremo opuesto del eje, probablemente asociados a menores ingresos per cápita.

En conjunto, los dos primeros componentes logran sintetizar una proporción relevante de la variabilidad del conjunto de datos, diferenciando principalmente entre hogares de mayor y menor nivel socioeconómico.

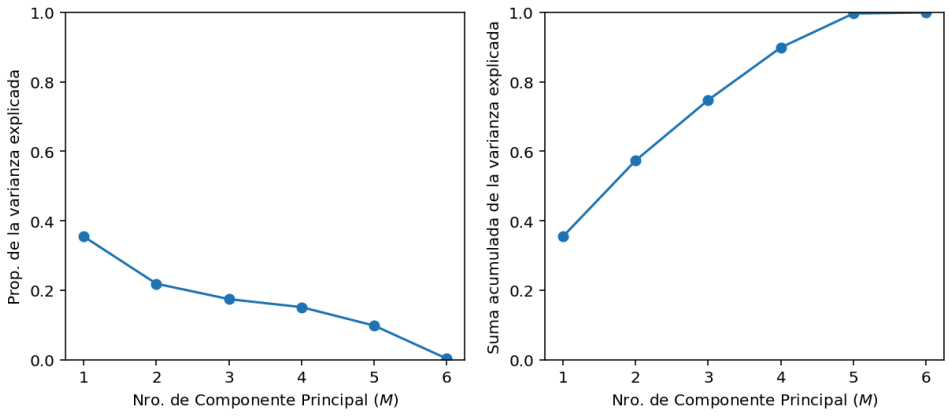



*4. Proporción de la varianza*

El gráfico de la izquierda muestra la proporción de varianza explicada por cada componente principal. En el eje horizontal se representan los componentes (CP1, CP2, etc.) y en el eje vertical la proporción de varianza que explica cada uno sobre el total. Se observa que el primer componente principal concentra alrededor del 35% de la varianza total, mientras que el segundo explica aproximadamente un 22%.

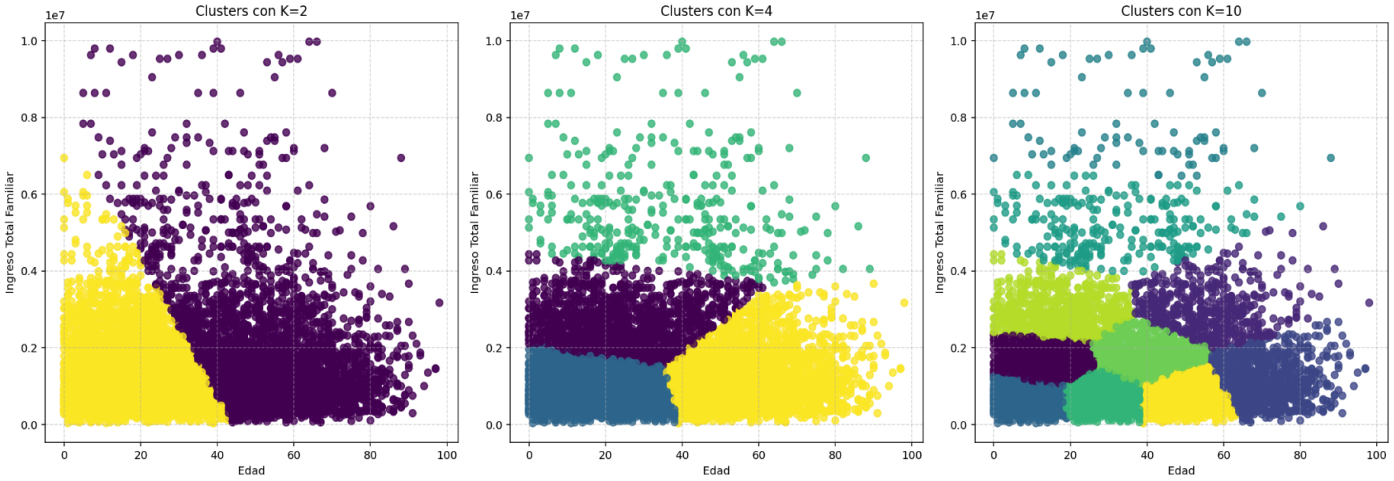
En conjunto, los dos primeros componentes capturan cerca del 58% de la variabilidad de los datos, lo que indica que contienen la mayor parte de la información relevante. A partir del tercer componente, la proporción de varianza explicada disminuye progresivamente, lo que sugiere que los componentes adicionales aportan menor información marginal.

El gráfico de la derecha muestra la varianza acumulada a medida que se incorporan nuevos componentes. Con los dos primeros se alcanza cerca del 60% de la varianza total, y con cuatro componentes alrededor del 90%. Esto indica que la reducción a dos o tres componentes resulta adecuada para representar la estructura principal de los datos sin perder información sustancial.



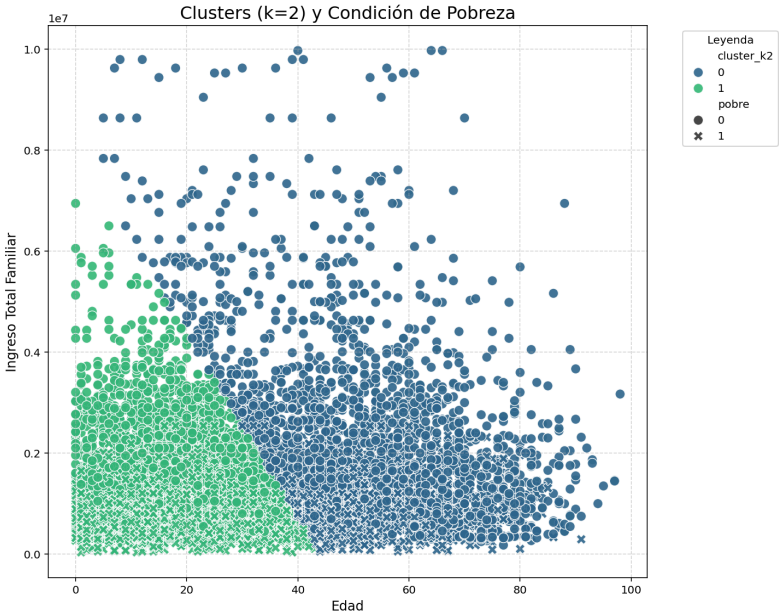
*5. Clusters*

*a.*




Se aplicó el algoritmo k-medias con k=2, 4 y 10 sobre las variables edad e ingreso total familiar.Con k=2, el algoritmo separa principalmente a la población según la edad, con un punto de quiebre alrededor de los 20-30 años. No se observa una división clara por nivel de ingreso, ya que ambos grupos presentan valores similares en promedio. Esto indica que la variable edad domina la clasificación. Al comparar con la condición de pobreza, se confirma que k=2 no distingue correctamente entre pobres y no pobres, sino entre grupos etarios.

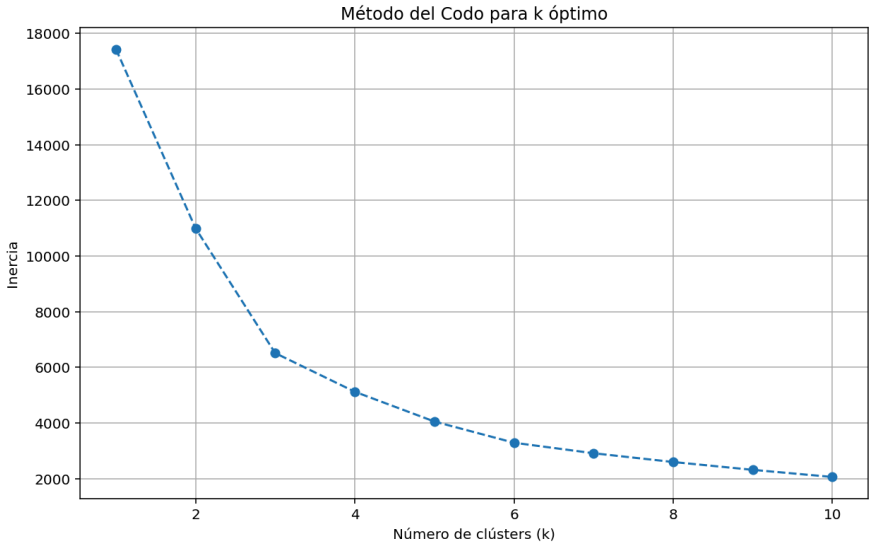
Con k=4 y k=10, el patrón se mantiene: los clusters se refinan en tramos etarios más específicos y de diferente nivel socioeconómico aportando un análisis mas detallado.

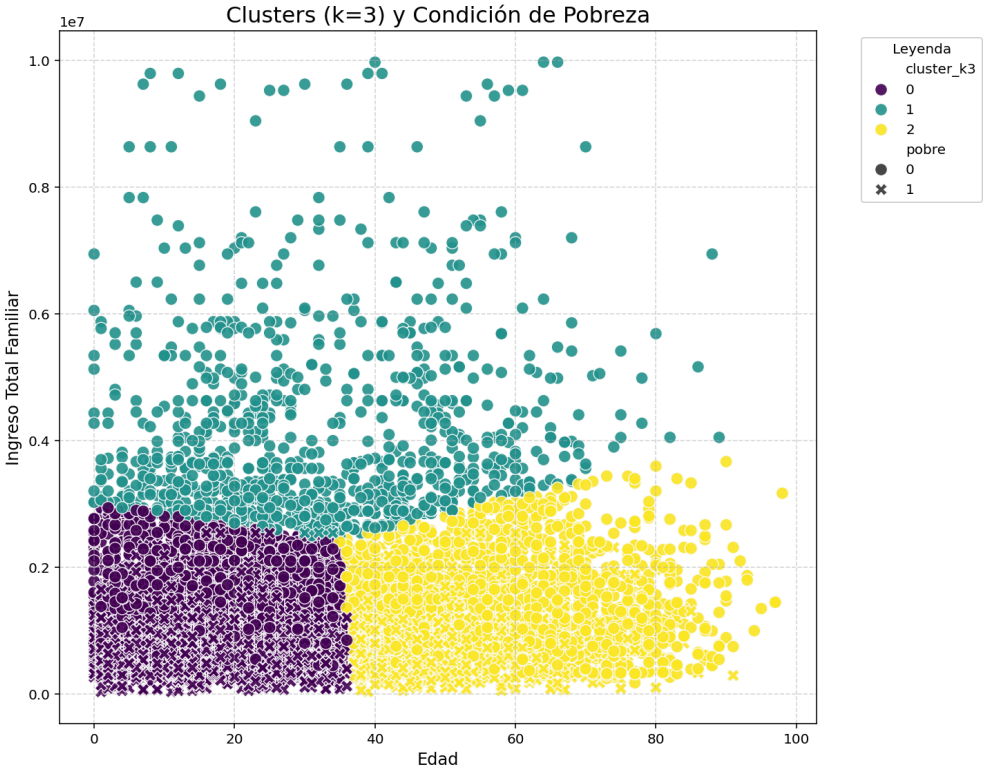


*b.*

Utilizando la medida de disimilitud de inertia, y graficando determinamos mediante el método del codo que 3 clusters son óptimos para el análisis ya que luego la caída observada en la inertia es muy leve.

Observando el gráfico de clustering con k=3 vemos que los dos clusters inferiores reflejan individuos en condición de pobreza mientras que el cluster superior agrupa individuos de mayor nivel socioeconómico. Sin embargo la delimitación no es exacta, con varios casos de no pobreza en los clusters bajos.



*6. Cluster jerárquico*

Un dendrograma es un diagrama en forma de árbol que ilustra el orden y la distancia a la que se combinan los grupos (clusters). El eje vertical ("Distancia") indica la disimilitud entre los grupos: cuanto más alta es la unión, mayor es la diferencia.

Este dendrograma truncado muestra la convergencia de cinco grupos finales hacia la raíz del árbol. Se observa que los clusters más pequeños, (3) y (24), se unieron a una baja distancia (similitud alta), y luego se fusionaron con el grupo más grande, (8688), a una distancia media. No obstante, la fusión final de los clusters restantes (4335 y 1573) con el mega-cluster ocurre a una distancia muy alta (cercana a 0.95). Esta gran altura de la última unión es la clave: sugiere una marcada disimilitud entre estos grandes grupos, lo que indica que una división natural del conjunto de datos en tres clusters podría ser la más apropiada para diferenciar segmentos de la población de la EPH basados en edad e ingreso.