Trabajo Práctico 3 'Big Data and Machine Learning'

Integrantes: Andrea Duette, Uriel Masserdotti, Paulo Gonzalez

Parte 1: Kernels e Histogramas

I

Con la variable "edad2" creada, elaboramos dos gráficos para ilustrar la distribución de la población según su estado de desocupación (ocupado o desocupado) y su edad. La 'Figura 1-1' muestra un histograma de las edades de los encuestados de la E.P.H. para 2.004 y 2.024, viéndose claramente una mayoría que tiene entre 20 y 50 años de edad, con un promedio de edad mayor en el 2.024. En general, se entrevistaron a más personas de distintas edades en 2.024.

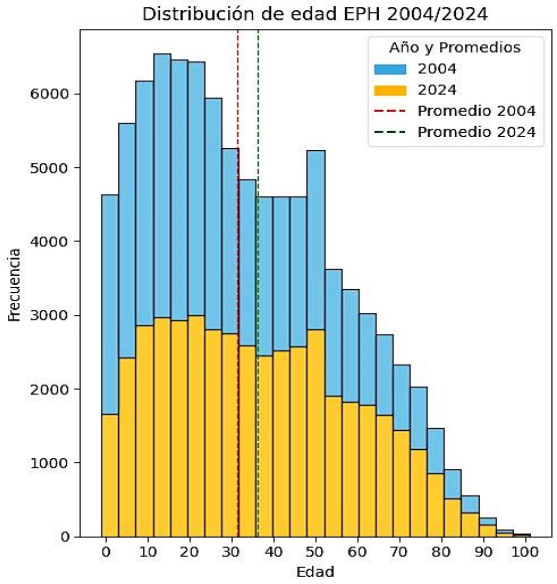

Figura 1-1: Histograma de la variable "edad2"

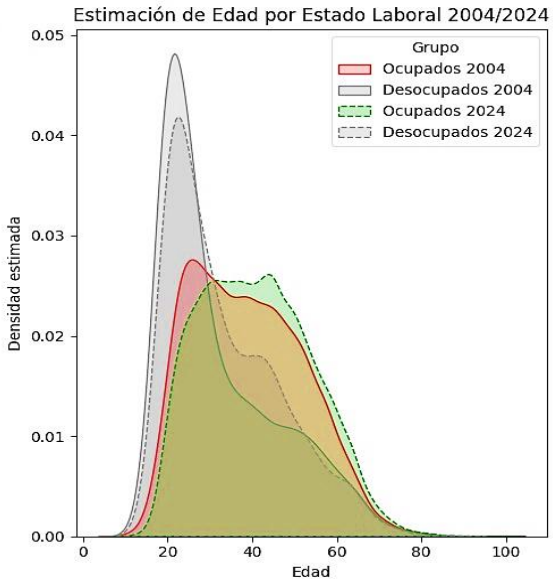

Figura 1-2: Distribución de kernels para los ocupados y desocupados.

En la 'Figura 1-2' se presentan las distribuciones de kernel de edad para ocupados y desocupados de los mismos años. Lo más llamativo resulta en el pico de desocupación que se alcanza entre los 20 y los 25 para ambos años (mayor sesgo a edades menores). Se sugiere un mayor nivel de desocupación asociado a edades más jóvenes y cierta estabilidad en los ocupados (predominantemente en aquellos entre los 30 y los 50 años de edad).

II

Creada la variable "educ" (que recolecta los años de educación de los encuestados), se ve un promedio de casi 7,94 años de educación, lo que indica una población encuestada que, en promedio, no alcanza el secundario completo y, además, se da un valor de mediana muy similar, sugiriendo una distribución con bajo sesgo y/o volatilidad (también representado por un desvío estándar de 4,58), con valores mínimos de 0 años de educación (personas sin ninguna educación formal) hasta máximos de 22 años (personas con altos niveles académicos).

III

Para el análisis del salario semanal para ambos periodos decidimos tomarnos ciertas libertades. En primer lugar, para convertir los salarios del 2004 a precios de 2024, usamos el aumento del Salario Mínimo Vital y Móvil (desde marzo de 2004 hasta marzo de 2024). Esa comparativa resulta en un S.M.V.M. de $350 en 2004 y, en 2024, de $202.800.

En segundo lugar, para obtener el salario semanal, en lugar de dividir por 40 (en referencia a las 40 horas trabajadas por semana), multiplicamos por $\frac{5}{21{,}65}$ (los 5 días de la semana pedidos en el ejercicio en el numerador, y los 21,65 días laborables que tiene un mes en el denominador) el salario promedio obtenido para ambos años (a precios de 2024). El uso de esta forma de calcular el salario semanal descansa sobre el hecho de que calculamos el salario semanal como una proporción del salario mensual, asumiendo que este se distribuye uniformemente a lo largo de los días laborales (21,65). La fórmula genérica tiene la forma: $salario\ semanal = salario\ mensual * \frac{días\ laborales\ en\ una\ semana}{días\ laborales\ en\ un\ mes}$.

Como último comentario antes de pasar a los resultados obtenidos, decidimos, para la 'Figura 1-4', hacer una distribución de kernels que relacione el salario semanal promedio con las edades de los encuestados. Esto por la razón de que, los desocupados (como pedía el ejercicio) no poseen salario alguno y no vienen al caso para analizar variaciones en el ingreso.

Para concluir, los resultados (graficados en las 'Figuras 1-3 y 1-4') muestran un promedio de salarios semanales mayor en el 2004 ($69.943 en 2004 contra $66.373 en 2024). La mayoría se los ingresos, para ambos periodos, se encuentran por debajo de este promedio. Como se ve en la 'Figura 1-4', hay una fuerte concentración de salarios semanales de entre $30.000 y $70.000 entre jóvenes de 20 años y adultos de hasta 40. Hay cierta tendencia de adultos con más años de experiencia (presumiblemente profesionales) que tienen ingresos mayores.

Los salarios más altos se observan en una edad media, entre 35 y 55 años, posiblemente de puestos de mayor jerarquía en el sector privado.

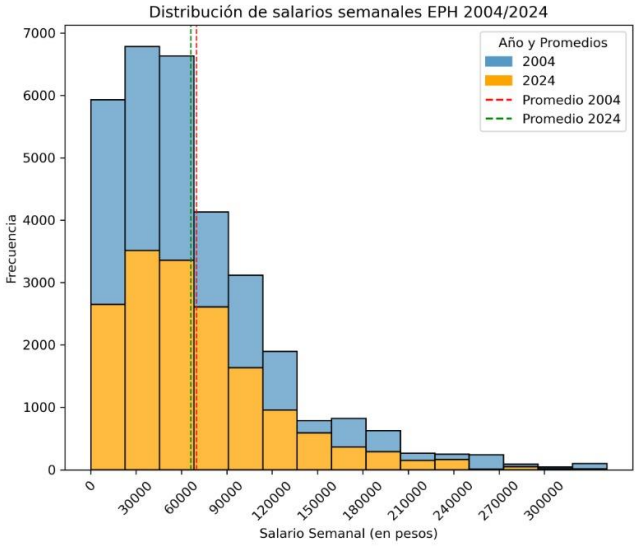

Figura 1-3: Histograma de la variable "salario_semanal".

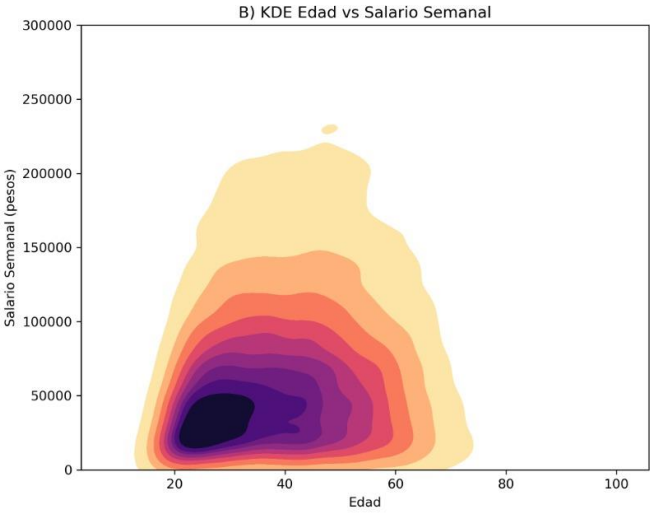

Figura 1-4: Distribución de kernels salario_semanal/edad.

IV

Análogo a lo expresado sobre la variable "educ", el promedio de horas trabajadas por los encuestados es de 20,6 horas trabajadas por semana, sugiriendo una tendencia hacia la subocupación laboral en los encuestados (incluso, al incluir desocupados en la muestra, estos pueden hacer tender a la baja el promedio). Sin embargo, si se ve la mediana y el desvío (8 y 23,8 respectivamente), se ve cierto sesgo y/o dispersión del promedio, indicando un alto rango de horas de trabajo semanal. Hay valores máximos elevados de 137 horas semanales (presumible un outlier) y, como es de esperarse, mínimos de 0 horas de trabajo.

Si se divide el análisis para cada año, la media de horas trabajadas en 2.024 es más del doble que la de 2.004 (35 y 14 horas respectivamente) con una desviación estándar levemente más alta para 2.004.

Si se vuelve a filtrar, pero esta vez, además de año, por franja de edad (entre 15 y 65 años) y por población económicamente activa, en 2.024 se trabaja levemente más que en 2.004: para el primer caso, se presenta un promedio de horas trabajadas de 35 horas con un desvío estándar de 19, mientras que para 2.004 el promedio es de 32 horas por semana y un desvío estándar de 24. En resumen: en 2.024 se trabaja más horas y jornadas más completas (con una distribución de muestras más concentradas en 40 horas semanales, con menos volatilidad).

V

Tratando la base de datos unificada para el Noreste Argentino, como se puede ver en la 'Tabla 1-1', se trabajó con una muestra de más de diez mil observaciones, con una mayoría de muestra para el 2.004. Se puede ver una buena calidad en la recolección de datos con solo un 0,1% de valores faltantes.

Se registraron 3.900 ocupados, con una mayoría para 2.024, y 342 desocupados con una mayoría para 2.004 (lo que se refleja en una diferencia positiva para ocupados del 18% y una diferencia negativa para desocupados del 41% entre ambos años).

	2.004	2.024	Total
Cantidad de Obs.	5.343	4.850	10.193
Cantidad de Obs. con "NAS" en variable "ESTADO"	5	6	11
Cantidad de Ocupados	1.789	2.111	3.900
Cantidad de desocupados	215	127	342
Cantidad de variables limpias y homogeneizadas	173	74	74

Tabla 1-1: Resumen de la base de datos unificada para la región del Noreste Argentino.

Parte 2: Métodos no supervisados

VI

Para empezar, en la 'Figura 2-1' se muestra una matriz de correlación para el Noreste Argentino de las variables con las que vamos a trabajar en esta segunda parte. Como se ve, por obvias razones, hay una fuerte razón entre la variable "edad" (CH06) y edad^2 (EDAD2), lo cual se verá también en el gráfico de análisis de componentes principales.

Se ve una relación negativa entre años de educación y edad, presumiblemente por el abandono o el cese del acceso a instituciones académicas con el paso del tiempo (menor escolaridad en adultos mayores). A su vez, hay una relación positiva entre el nivel de salario semanal y los años de educación (casi tan fuerte como horas trabajadas y salario).

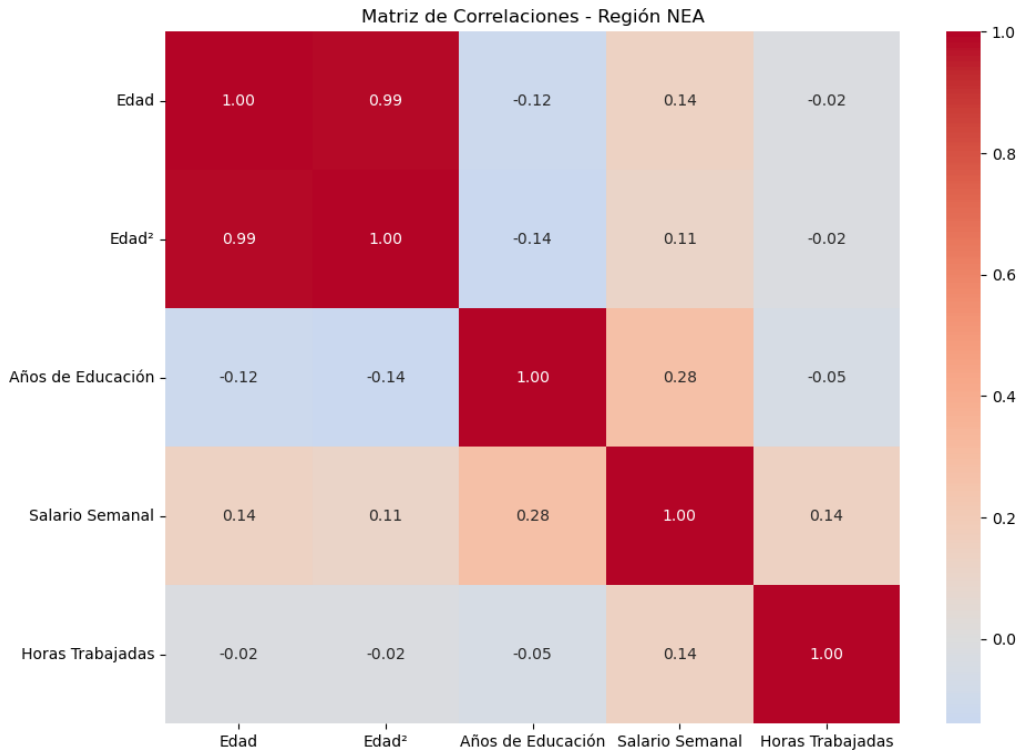

Figura 2-1: Matriz de Correlación de cinco predictores del Noreste Argentino.

VII

Siguiendo con el análisis de estos predictores, la 'Figura 2-2' presenta un gráfico de dispersión que muestra los índices (scores) del primer y segundo componente principal (PC1 y PC2) obtenidos al aplicar P.C.A. sobre las variables estandarizadas presentadas anteriormente. Las flechas de los loadings indican cómo el salario semanal, las horas trabajadas y el nivel de educación aportan principalmente al componente 2 (eje vertical), mientras que la edad y la edad al cuadrado (al estar casi perfectamente relacionadas entre sí) aportan fuertemente al componente 1 (eje horizontal).

En este sentido, el primer componente captura variables demográficas mientras que el segundo recopila datos laborales y educacionales.

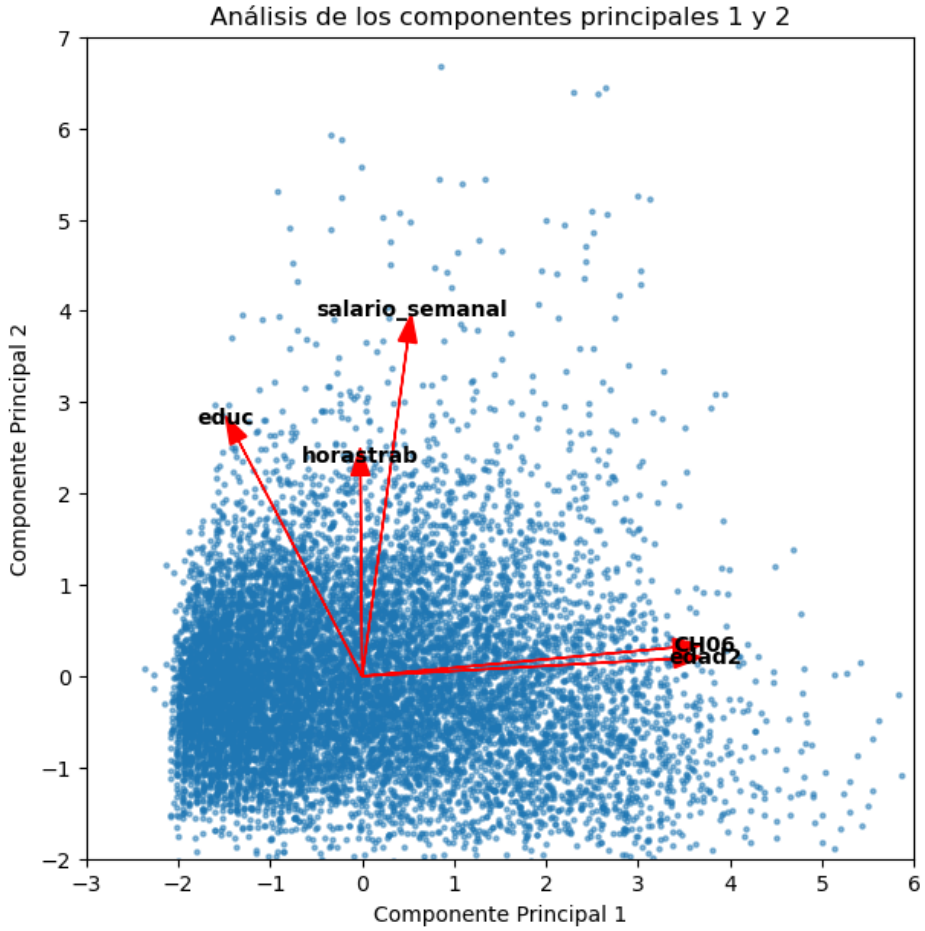

Figura 2-2: Gráfico de dispersión de los scores de los componentes principales.

VIII

De forma complementaria, en las 'Figura 2-3' y '2-4' se ve la varianza que aportan cada uno los componentes principales (en rojo) y la varianza acumulada de las variables estandarizadas (en verde). Así, se ve que entre el PC1 y el PC2 explican el 67% de la varianza total. Es de resaltar el hecho de que el quinto componente no explique casi nada (0,3% de la varianza explicada), pues se desprende del hecho de que la edad y la edad al cuadrado están muy interrelacionadas.

Para un análisis más completo puede llegar a ser útil incluir el tercer componente principal para así poder explicar el 86% de la varianza acumulada.

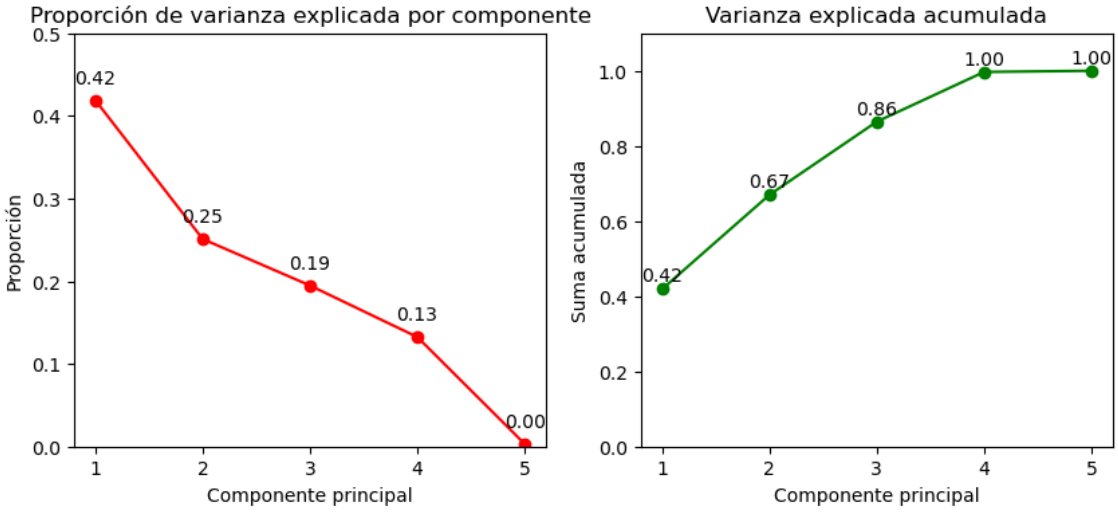

Figura 2-3: Gráfico de proporción de la varianza explicada por cada componente principal.

Figura 2-4: Gráfico de proporción de la varianza explicada acumulada.

Relacionado con el análisis de los componentes principales se desprende el análisis de clustering con k-medias para la misma región, ilustrado en las 'Figuras 2-5, 2-6, 2-7'.

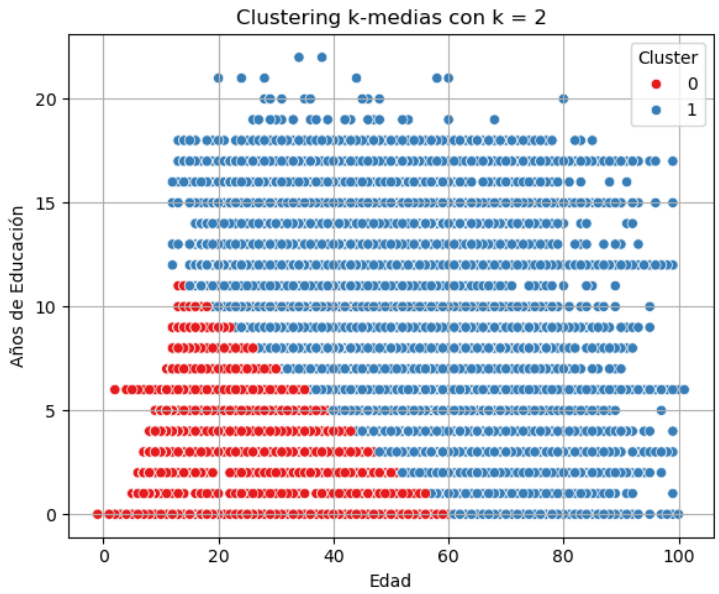

Figura 2-5: Clustering k-medias con K=2.

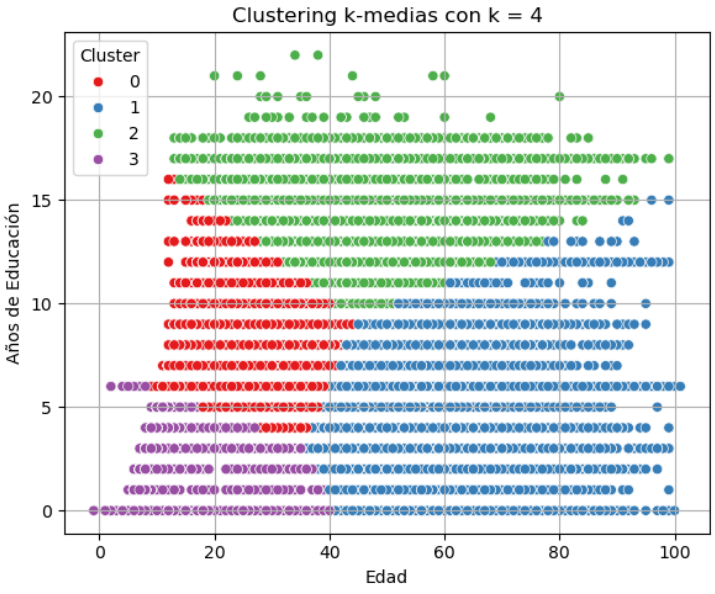

Figura 2-6: Clustering k-medias con K=4.

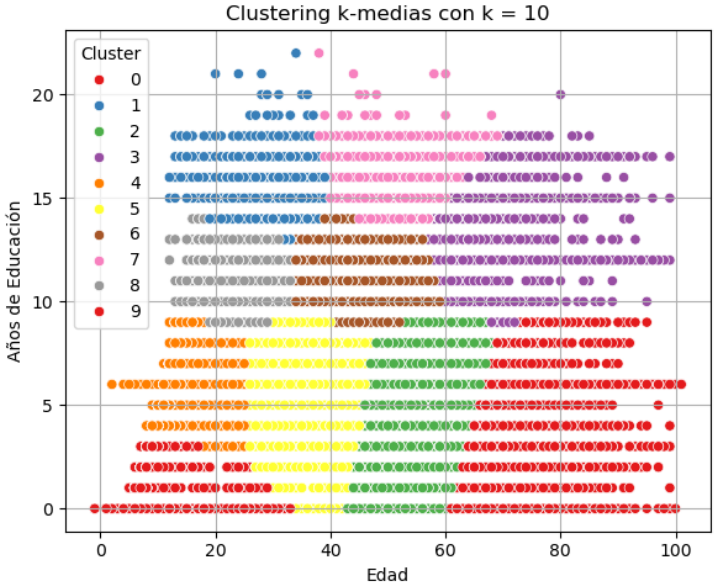

Figura 2-7: Clustering k-medias con K=10.

En la 'Figura 2-5' se ilustra el análisis de k-medias con un k=2. En este caso el "cluster 0" (rojo) agrupa a personas con valores bajos de la variable "nivel educativo", concentrado en adultos jóvenes. El "cluster 1" abarca un rango más amplio de edad y educación, sobre todo adultos con distintos niveles académicos.

En el caso de la 'Figura 2-6' se puede apreciar un k=4 con cluster algo más específicos. El "cluster 0" (rojo) incluye personas de 15 a 40 años con 6 a 13 años de educación. Este análisis es análogo a los otros cluster, con, por ejemplo, el "cluster 2" incluyendo a personas de 20 a 60 años con más de 13 años de educación (profesionales con estudios superiores o títulos de posgrado).

Por último, la 'Figura 2-7' es la más interesante, ya que muestra subgrupos de clusters más detallados. Puede que sea el clustering más útil, ya que permite segmentar más finamente el análisis y explorar distintos subgrupos específicos de la muestra. Por ejemplo, si bien los clusters 7, 3 y 0 agrupan personas con educación superior, una vez fragmentada esa agrupación, puede interpretarse como al cluster 7 representando jóvenes que cursan estudios superiores, el cluster 3 como adultos ya recibidos de una carrera terciaria y al cluster 0 como adultos con títulos de posgrado o varios títulos de grado con profesiones consolidadas.

IX

Ahora pasamos a aplicar k-means con K=2 para los scores, diferenciando a las personas ocupadas (en verde) de las desocupadas (en amarillo) en el gráfico de dispersión. A su vez, para diferenciar los clusters, en lugar de colores como en las anteriores figuras, se usaron formas (círculos para el cluster 0 y cruces para el cluster 1).

En la 'Figura 2-8' se ven dos clusters: el "cluster 0" agrupa a personas (mayoritariamente) más jóvenes y con menos educación. Mientras que el "cluster 1" agrupa a personas mayores y con más educación (en cierta forma similar a la 'Figura 2-5').

Si bien los desocupados están dispersos por ambos clusters, muestran una ligera mayor concentración en el cluster 0 (presumiblemente jóvenes aun cursando sus estudios).

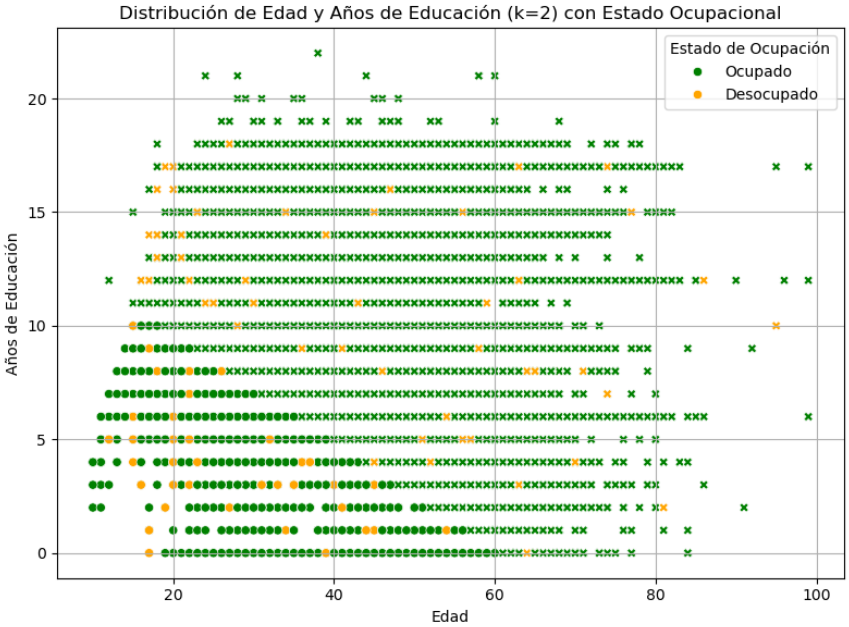

Figura 2-8: Clustering k-medias con K=2 para ocupados y desocupados.

X

Por último, para ilustrar otra forma en la que se agrupan las observaciones, se propone un dendograma de las variables elegidas (Figura 2-9). En el eje horizontal se muestran las observaciones agrupadas que corresponden a las filas de la base de datos (una vez aplicado el análisis de componentes principales). Mientras que en el eje vertical se mide la distancia de disimilitud entre las observaciones (clusters) a medida que se fusionan.

Se puede ver cómo, al inicio, cada observación es un cluster individual y que, a medida que se sube en el eje vertical, los clusters se fusionan gradualmente según su similitud. El corte sugerido de clusters se realizó a una altura donde se identifican aproximadamente 30 clusters antes de que se fusionen en un solo grupo.

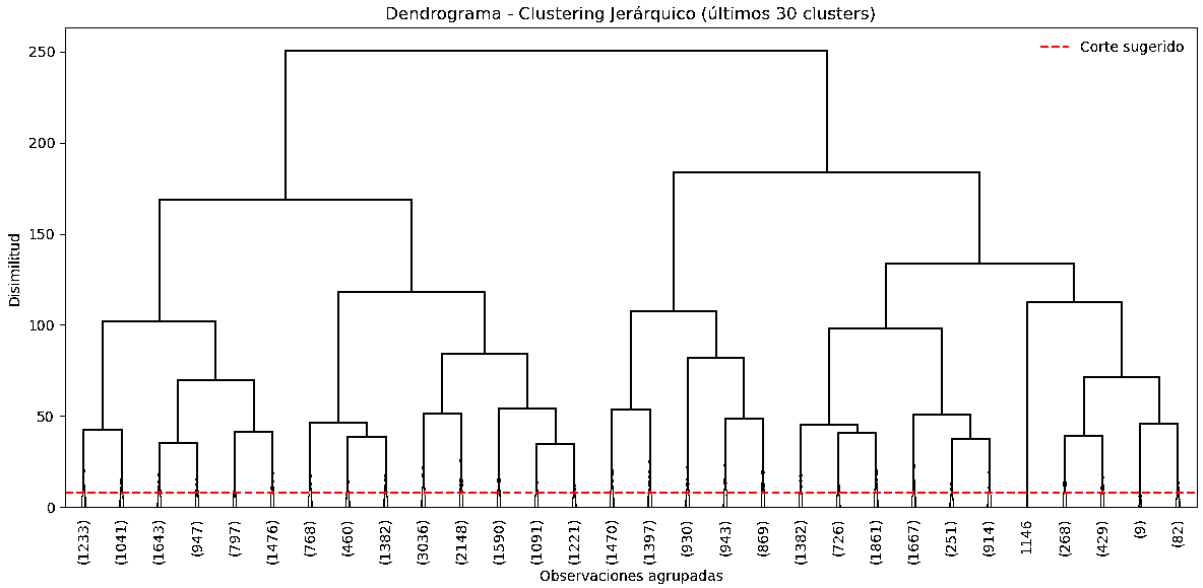

Figura 2-9: Dendograma jerarquico de clusters aplicado a los componentes principales.