UNIVERSIDAD DE BUENOS AIRES - 2025

BIG DATA & MACHINE LEARNING

TRABAJO PRÁCTICO N° 3: HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

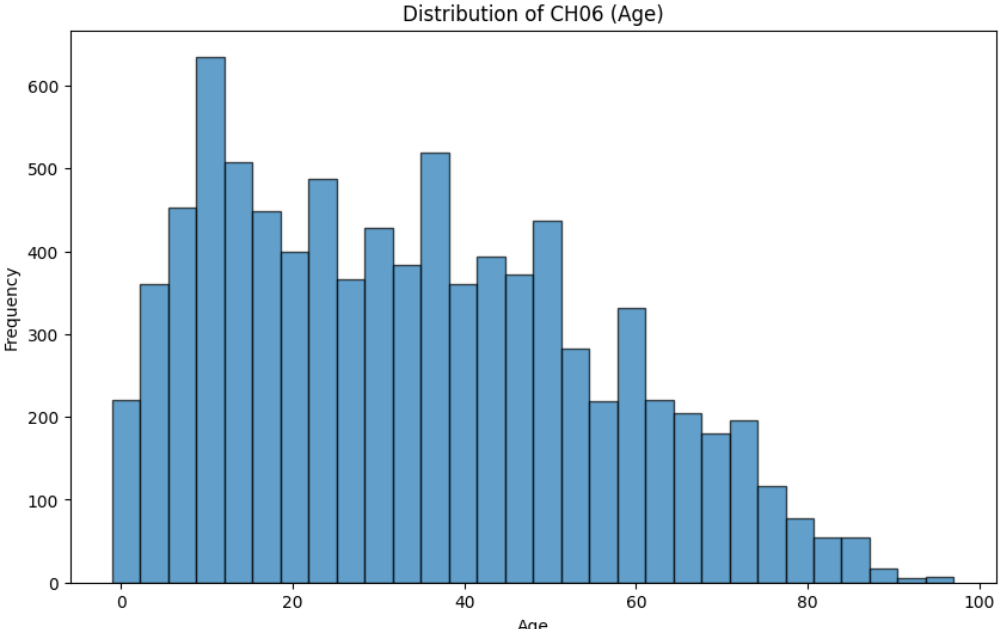
Fecha de entrega: Mayo 13 de marzo a las 13:00 hs.

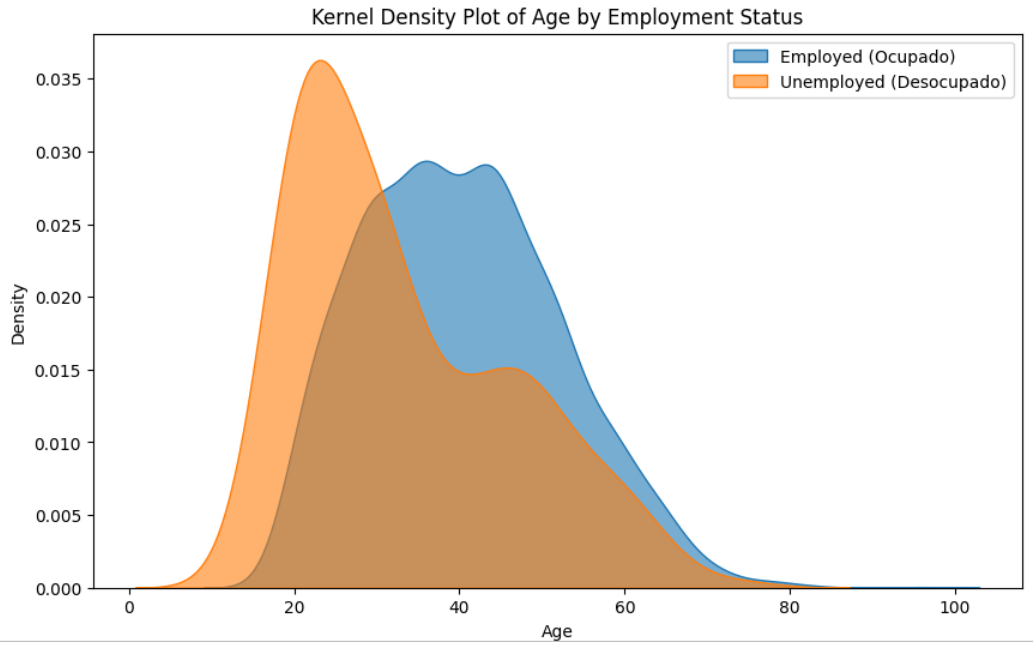
Rosario Llamazares

Dalma Galvan

Shanaia San Roman

En los dos paneles correspondientes a la región Patagónica se observa que la distribución de edades se concentra principalmente en la población adulta, en especial entre los 25 y 50 años. Al eliminar los valores extremos (menores de 1 año y mayores de 97), se logra una representación más precisa del grupo en edad económicamente activa. En el gráfico de densidad, se nota que las personas ocupadas tienden a tener una edad media ligeramente mayor que las desocupadas. Esto podría indicar que la inserción laboral mejora con la experiencia o la edad, aunque se necesitaría un análisis más profundo para confirmarlo.





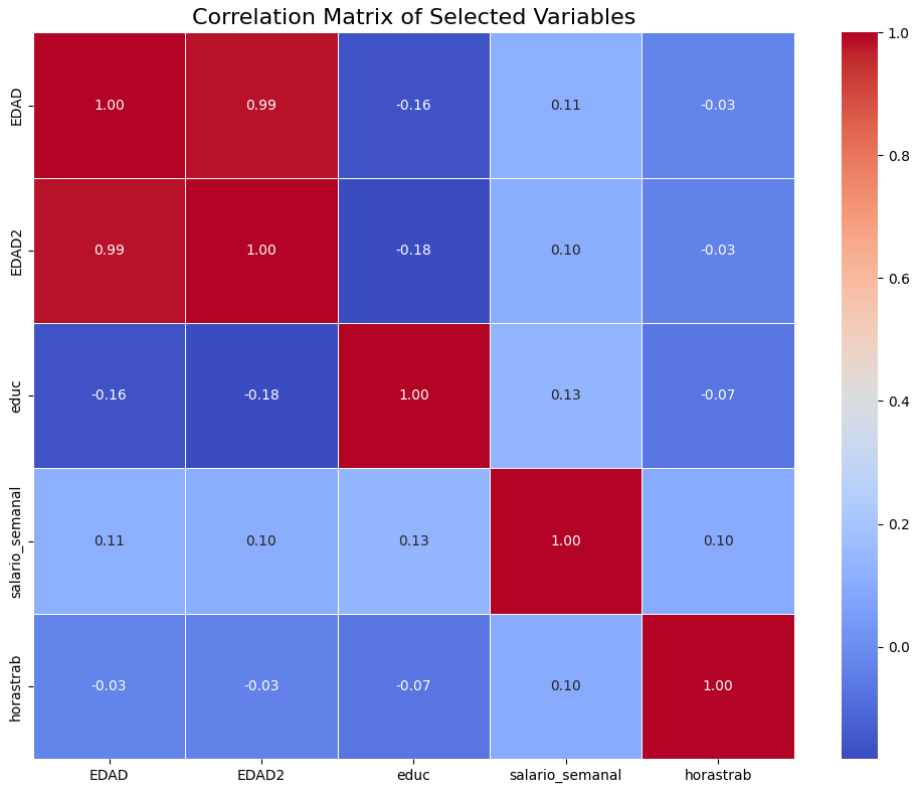
La variable educ fue construida a

partir de tres variables del cuestionario: CH12, que indica el nivel educativo alcanzado (como primario, secundario, universitario, etc.); CH13, que señala si ese nivel fue completado; y CH14, que informa el último año aprobado dentro de ese nivel. A partir de esta información, se estiman los años de educación efectivos de cada persona. Por ejemplo, si una persona declaró tener nivel secundario incompleto y haber aprobado hasta segundo año, el código le asigna 9 años de educación (7 de primaria más 2 de secundaria). Esta variable permite analizar el capital educativo de manera más precisa, lo cual es importante para estudiar su relación con el empleo, los ingresos y otras variables socioeconómicas.

La distribución de los **salarios semanales** muestra una gran **asimetría positiva**, con un salario promedio de 6,809.69 pesos, pero una **alta variabilidad** (desviación estándar de 9,601.97), lo que indica que hay una gran concentración de salarios bajos. El **primer cuartil** está en 23.5 pesos y la **mediana** en 5,000 pesos, lo que significa que la mayoría de los trabajadores ganan menos que el promedio. La presencia de valores extremos, como un salario máximo de 250,000 pesos, eleva el promedio, pero no refleja la realidad de la mayoría de los trabajadores.

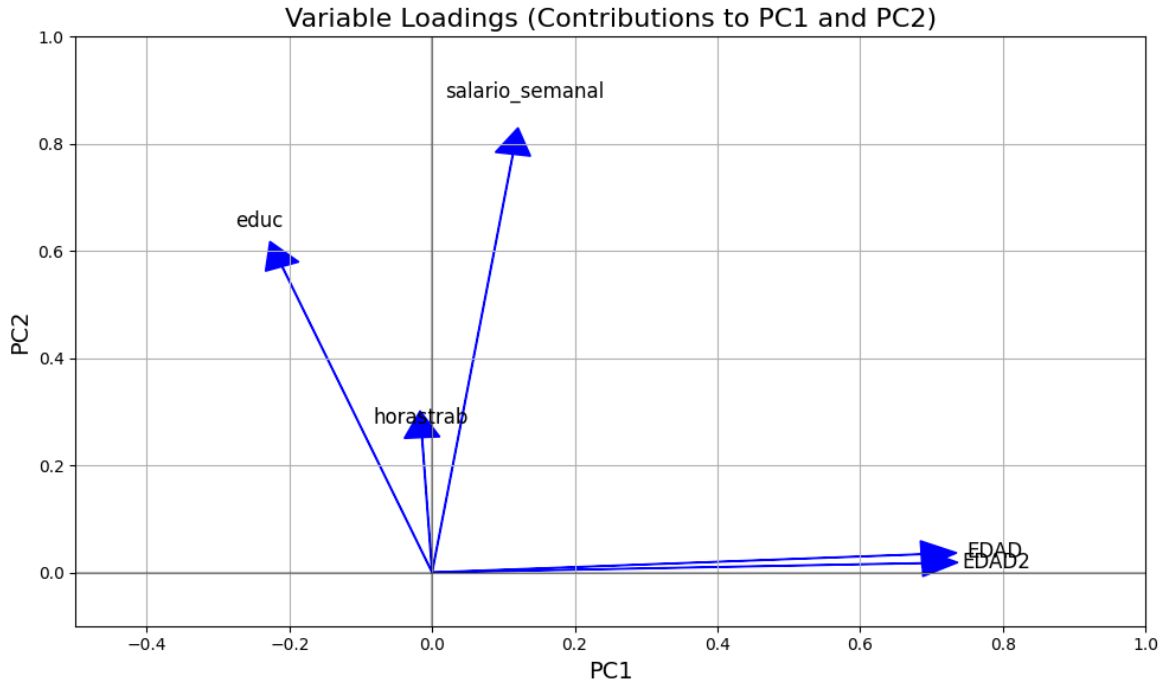
La variable **horastrab** muestra un promedio de **37.42 horas** trabajadas por semana, lo que se ajusta a una jornada laboral estándar, pero con una **gran variabilidad** (desviación estándar de 19.69 horas). La presencia de un **mínimo de 0 horas** sugiere que algunas personas no están trabajando, mientras que el **máximo de 126 horas** podría reflejar jornadas extremadamente largas. Esto indica que, aunque la mayoría de las personas trabaja alrededor de 40 horas, hay una gran diferencia en las horas de trabajo entre los individuos.

En el gráfico se ve que la variable **EDAD** y **EDAD2** están casi perfectamente correlacionadas (0.99), lo cual tiene sentido porque una es el cuadrado de la otra. La educación (**educ**) tiene una correlación leve y negativa con la edad, y una relación muy débil con el salario y las horas trabajadas. En general, las correlaciones entre las variables económicas son bajas, lo que sugiere que no hay una relación lineal fuerte entre ellas.

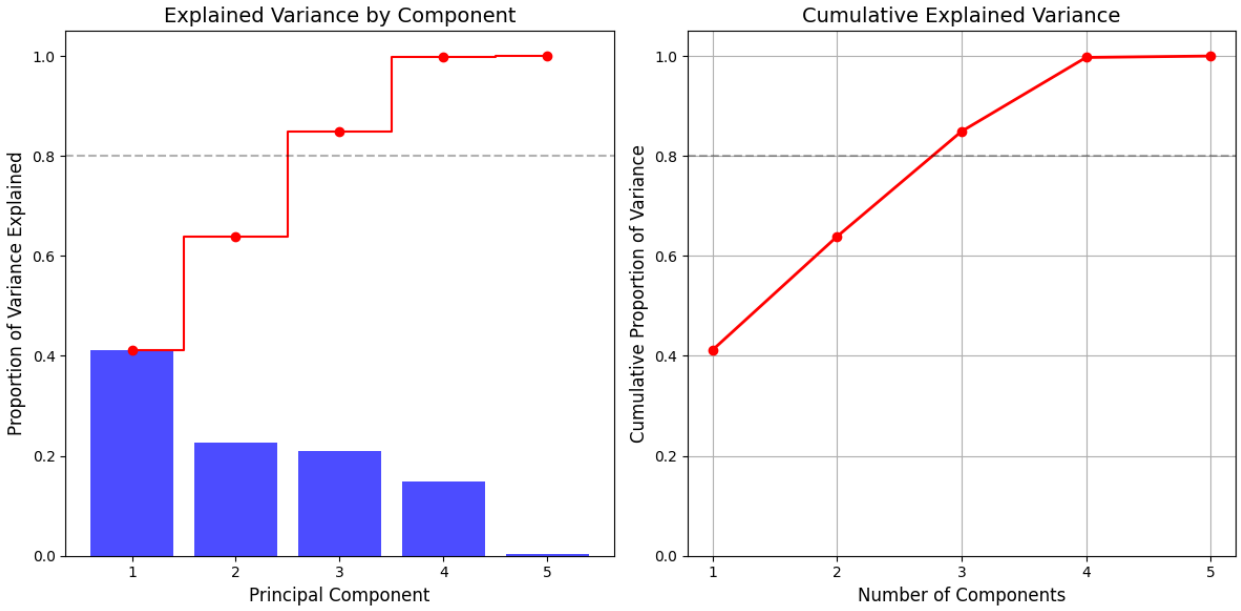


Las cargas de los componentes principales muestran que la edad y su cuadrado (EDAD y EDAD2) influyen principalmente en la primera componente (PC1), lo que sugiere que la edad es una variable clave en este análisis. Por otro lado, el salario semanal y la educación tienen una gran influencia en la segunda componente (PC2), lo que indica que estos factores están más relacionados entre sí. En resumen, el análisis revela cómo las variables económicas y demográficas se agrupan y contribuyen de manera diferente a las componentes principales, ayudando a reducir la complejidad del conjunto de datos.

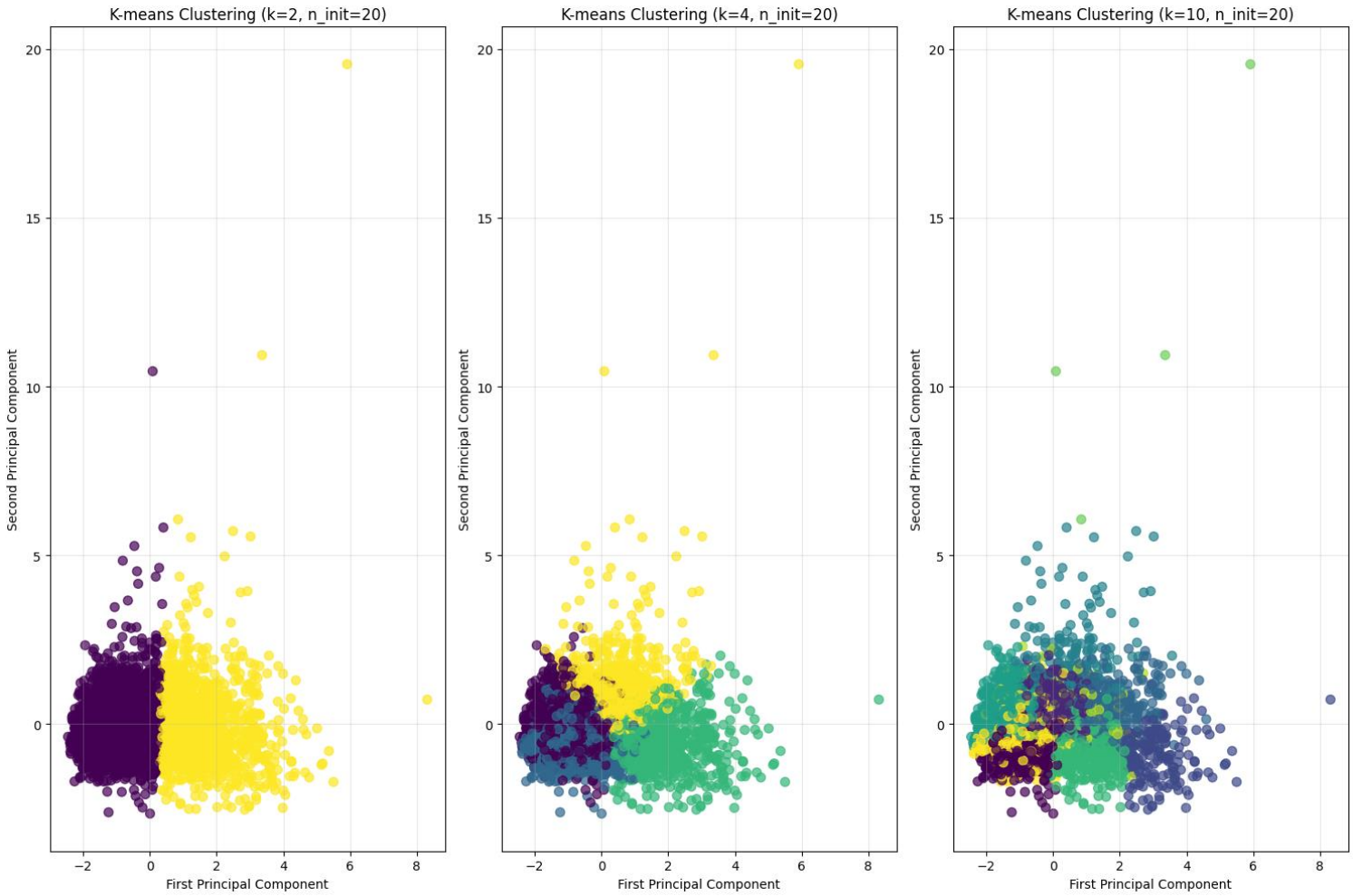
En este gráfico de componentes principales se ve que **EDAD** y **EDAD2** explican casi todo el eje PC1, o sea que son las variables que más peso tienen en esa dimensión. Por otro lado, **salario_semanal** y **educ** tienen más carga en PC2, lo que indica que ayudan a diferenciar observaciones en otra dirección. Las flechas cortas como la de **horastrab** muestran que esa variable tiene poca influencia en estos dos componentes principales.



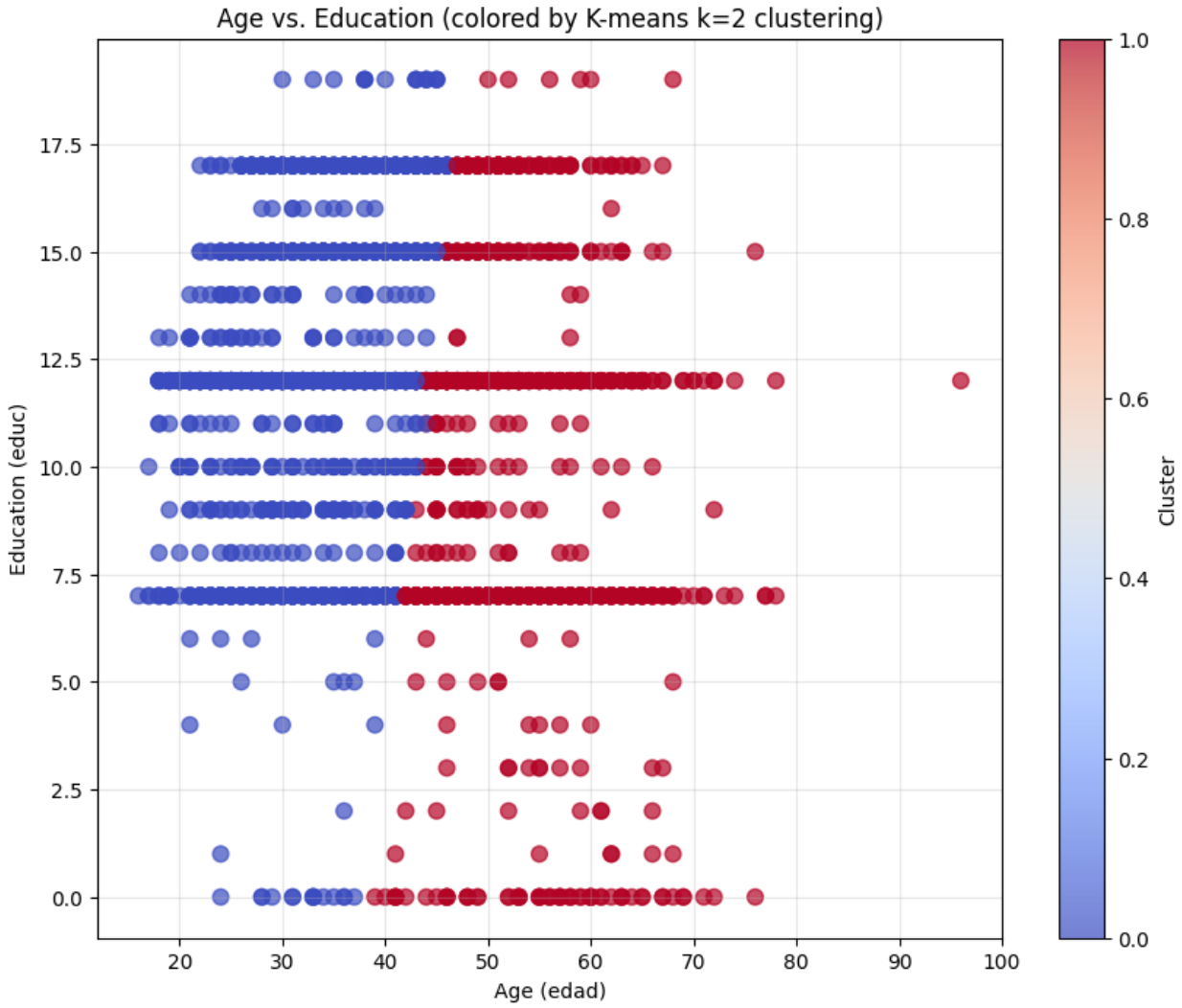
En el gráfico se ve que los **primeros tres componentes** explican más del **85% de la varianza total**, lo que indica que ya capturan la mayor parte de la info del dataset. El **primer componente** por sí solo explica alrededor del **40%**, lo cual es bastante. A partir del cuarto componente, el aporte es muy chico, así que no suma mucho incluir más de tres.



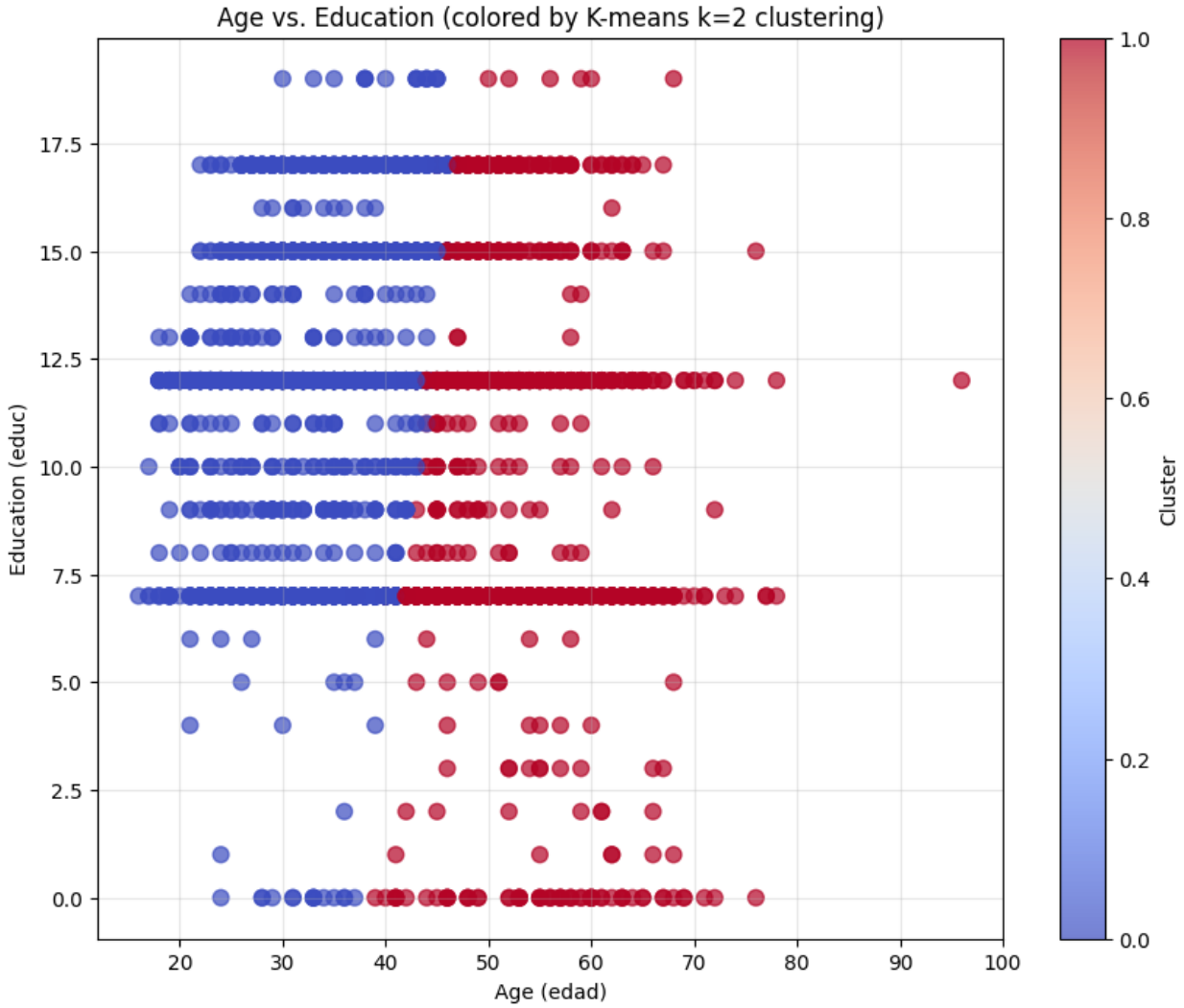
Interpretando los gráficos con una mirada más económica, los mismos podrían representar distintos grupos de trabajadores con características similares, como niveles de salario, educación y edad. Por ejemplo, al usar k=4 podríamos estar identificando perfiles como jóvenes con baja educación y salario bajo, o personas mayores con más experiencia y salario más alto. Esto nos sirve para entender mejor la segmentación del mercado laboral y cómo se relacionan las variables entre sí.



Se puede ver que los grupos están separados principalmente por edad: el **azul** concentra personas más jóvenes, mientras que el **rojo** incluye mayormente a personas mayores, muchas con baja educación. Sin embargo, como el algoritmo **no utiliza directamente la variable de ocupación**, no podemos asegurar que estos graficos representen fielmente a ocupados y desocupados; es posible que haya mezcla dentro de cada grupo.



Un dendograma es un gráfico que usamos para ver cómo se agrupan los datos en base a su similitud, como cuando tratamos de clasificar personas según sus características. En este caso, los datos se van juntando en ramas, y si cortamos el árbol a cierta altura, podemos decidir cuántos grupos hay. En economía, esto puede servir para segmentar población o mercados sin tener que asumir cosas de entrada, algo bastante útil para el análisis exploratorio.


Age vs. Education (colored by K-means k=2 clustering)
Education (educ)
Age (edad)
Cluster
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20
30
40
50
60
70
80
90
100
0.0
0.2
0.4
0.6
0.8
1.0