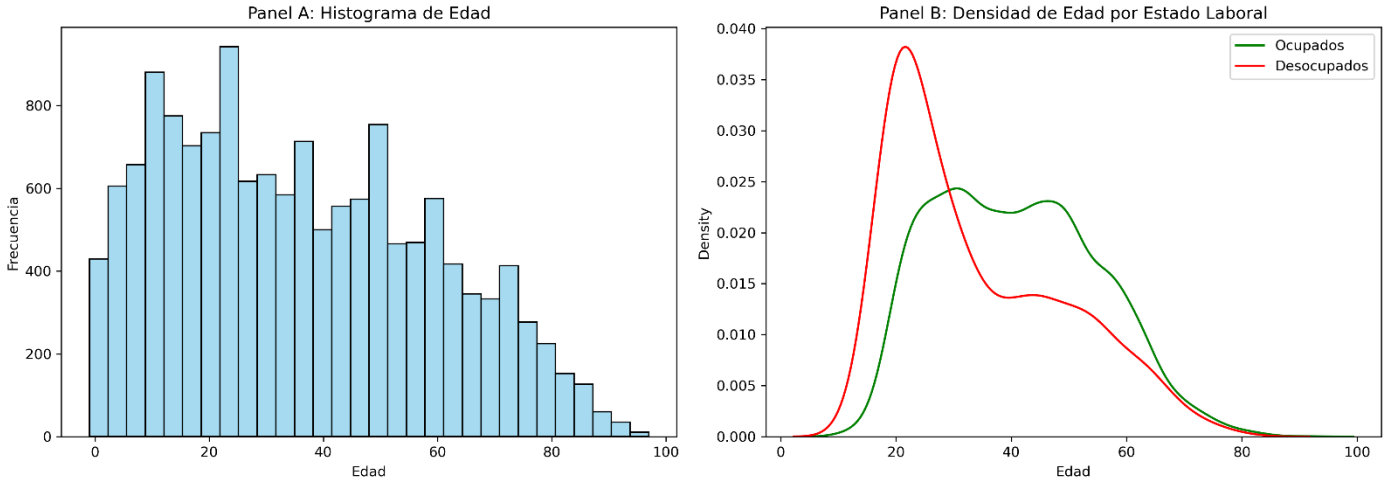
BIG DATA & MACHINE LEARNING

TRABAJO PRÁCTICO N° 3

HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

PARTE I:

1)



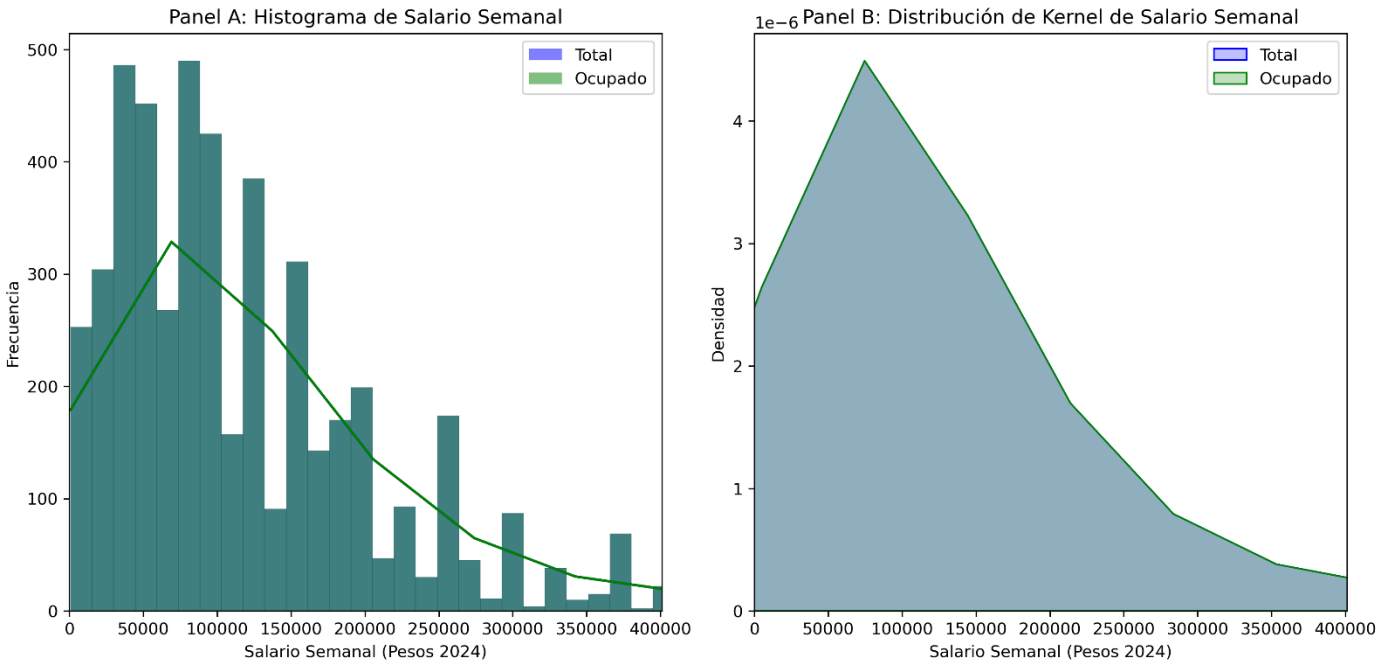
Los gráficos muestran claramente la distribución etaria de la muestra. En el **Panel A**, se observa una concentración de individuos entre los 10 y 30 años, con una caída progresiva hacia edades mayores. El **Panel B** revela que los **desocupados se concentran en edades más jóvenes**, especialmente alrededor de los 20 años, mientras que los **ocupados tienen una distribución más extendida** hasta los 60 años. Esto sugiere una inserción laboral más limitada entre los más jóvenes.

2)

	Media	Desviación estándar	Mínimo	Mediana	Máximo
Años de Educación	11,03069	3,362596	7	12	20

La media de 11.03 años de educación indica que, en promedio, las personas tienen algo más de 11 años de escolaridad. La desviación estándar de 3.36 refleja una variabilidad moderada en los años de educación. El mínimo de 7 años y el máximo de 20 años muestran una amplia gama de niveles educativos, mientras que la mediana de 12 años sugiere que la mayoría de las personas tiene alrededor de la secundaria completa.

3)



En el **Panel A** se observa que la mayor parte de los asalariados semanales gana entre $50.000 y $100.000 (valores ajustados a 2024), con una clara concentración en los tramos más bajos. El **Panel B** refuerza esta observación mostrando una distribución sesgada a la derecha, lo que indica la existencia de una minoría con salarios significativamente más altos. La superposición de curvas sugiere que casi todos los asalariados en la muestra corresponden a personas ocupadas. En conjunto, los gráficos revelan una alta desigualdad en la distribución del ingreso semanal.

4)

	Media	Desviación estándar	Mínimo	Mediana	Máximo
Horas Trabajadas	36.78	21.58	0	40	133

La media de 36.78 horas sugiere que, en promedio, se trabaja casi 37 horas. La gran desviación estándar (21.58) indica una gran variabilidad. El mínimo de 0 horas y el máximo de 133 horas destacan casos extremos, mientras que la mediana de 40 horas sugiere una tendencia hacia jornadas largas.

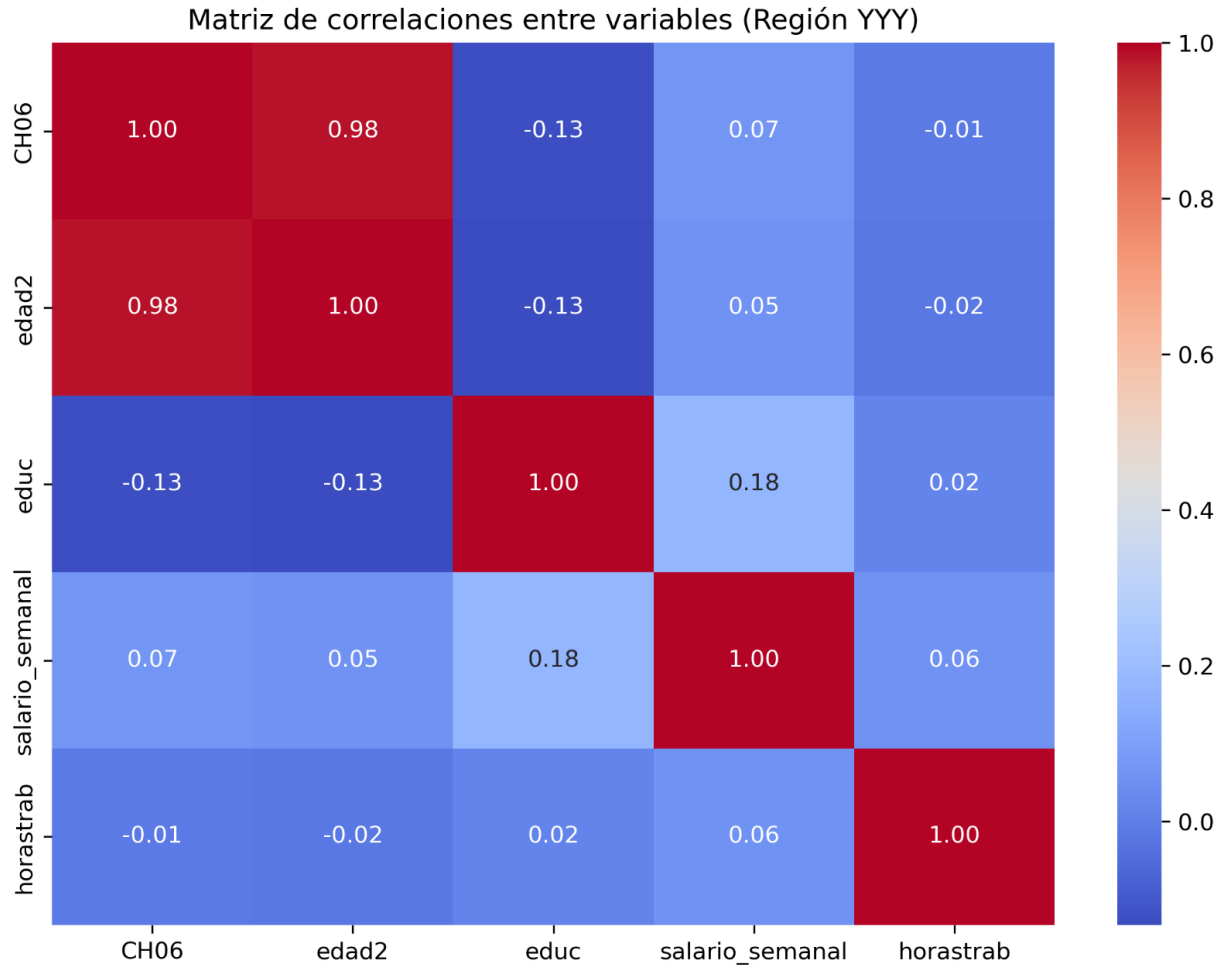
5)

	2004	2024	Total
Cantidad observaciones	7647	7051	14698
Cantidad de observaciones con NaNs en "ESTADO"	0	0	0
Cantidad de Ocupados	3079	3224	6303
Cantidad de Desocupados	528	311	839
Cantidad de variables limpias y homogeneizadas	187	187	187

La base de datos cuenta con un total de **14.698 observaciones**, distribuidas en **7.647 registros para el año 2004** y **7.051 para el año 2024**. En cuanto al mercado laboral, se observa un ligero **aumento en la cantidad de ocupados**, pasando de 3.079 en 2004 a 3.224 en 2024, mientras que la cantidad de desocupados **disminuyó** notablemente, de 528 a 311 casos. Finalmente, la cantidad de **variables limpias y homogeneizadas es consistente** entre ambos años, con un total de 187 variables, lo que garantiza comparabilidad y solidez para próximos análisis.

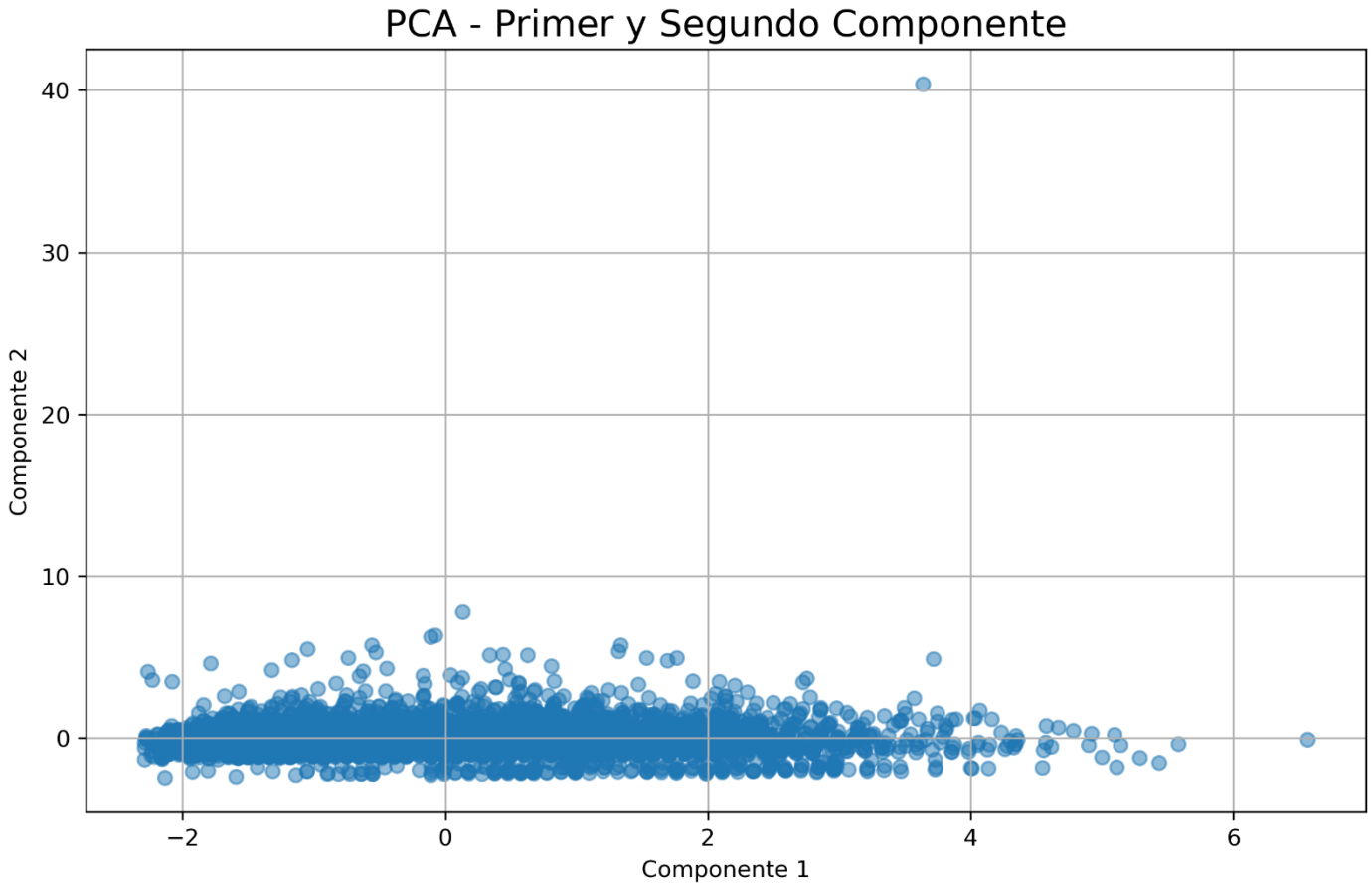
Parte II:

1)



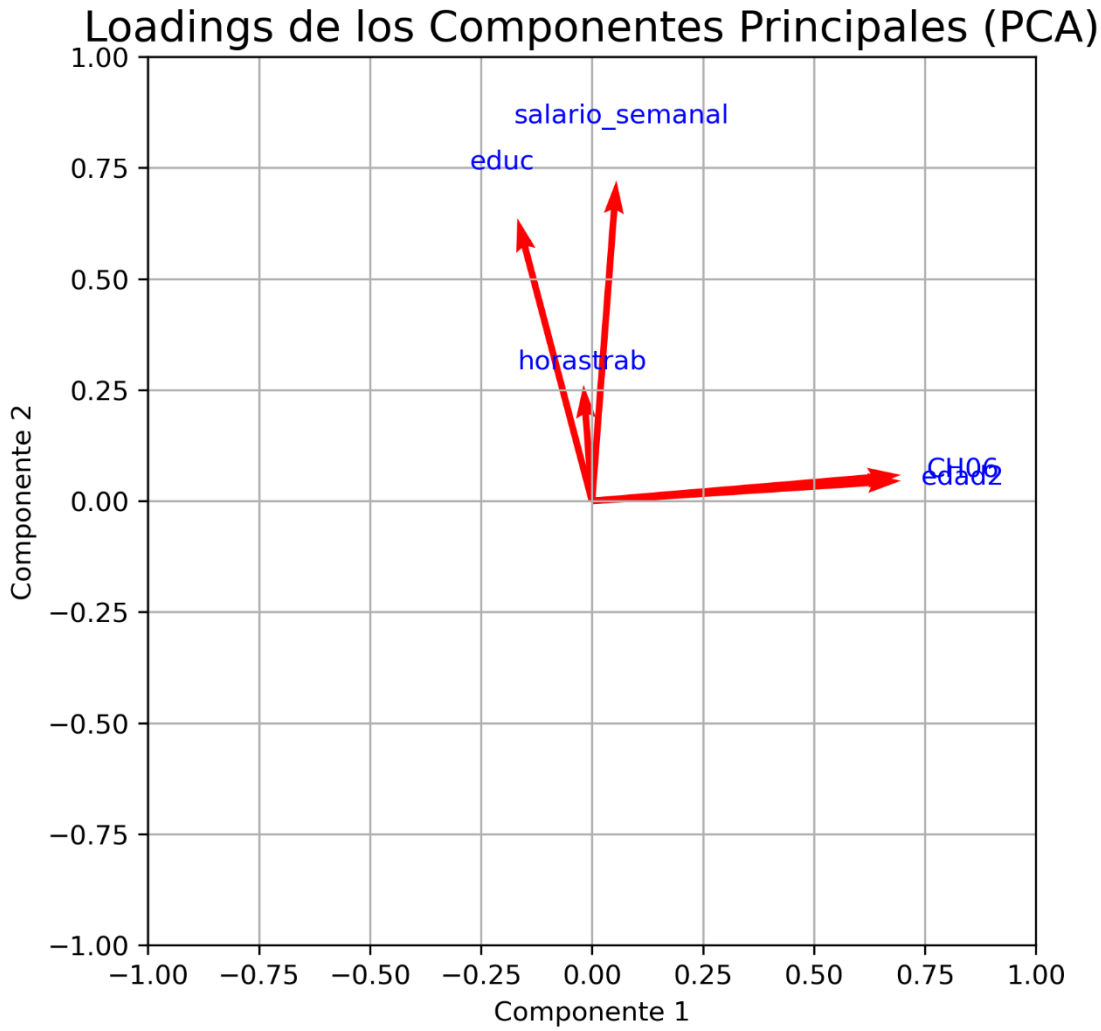
La matriz muestra una **fuerte correlación positiva entre CH06 (edad) y edad2 (edad al cuadrado)**, lo cual es esperable dado que una es función de la otra. La **educación tiene una correlación débil y positiva con el salario semanal** (0.18), lo que sugiere una relación directa pero no determinante entre mayor nivel educativo e ingresos. Por otro lado, las **horas trabajadas no están fuertemente correlacionadas** ni con el salario ni con la edad. En general, las relaciones entre variables son bajas, lo que sugiere cierta independencia entre dimensiones como educación, edad e ingreso dentro de esta región.

2)



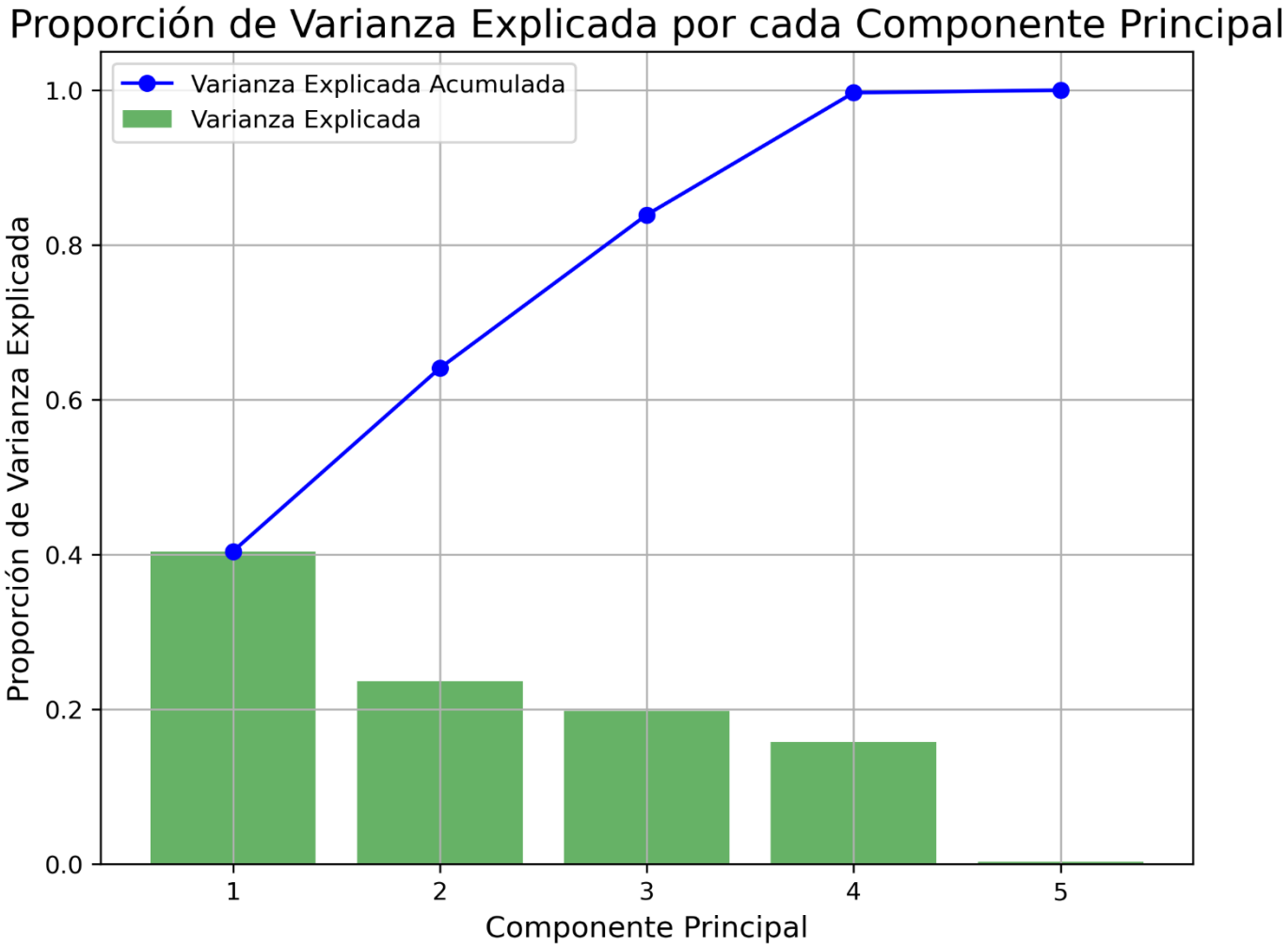
El gráfico de componentes principales muestra cómo los datos se distribuyen en el espacio reducido de las dos primeras dimensiones. Se observa una mayor dispersión a lo largo del primer componente, lo que sugiere que este concentra la mayor parte de la variabilidad explicada. La mayoría de los puntos están agrupados cerca del eje horizontal, indicando que el segundo componente aporta menos variabilidad. También se destacan algunos outliers que podrían influir en el análisis si no se tratan adecuadamente.

3)



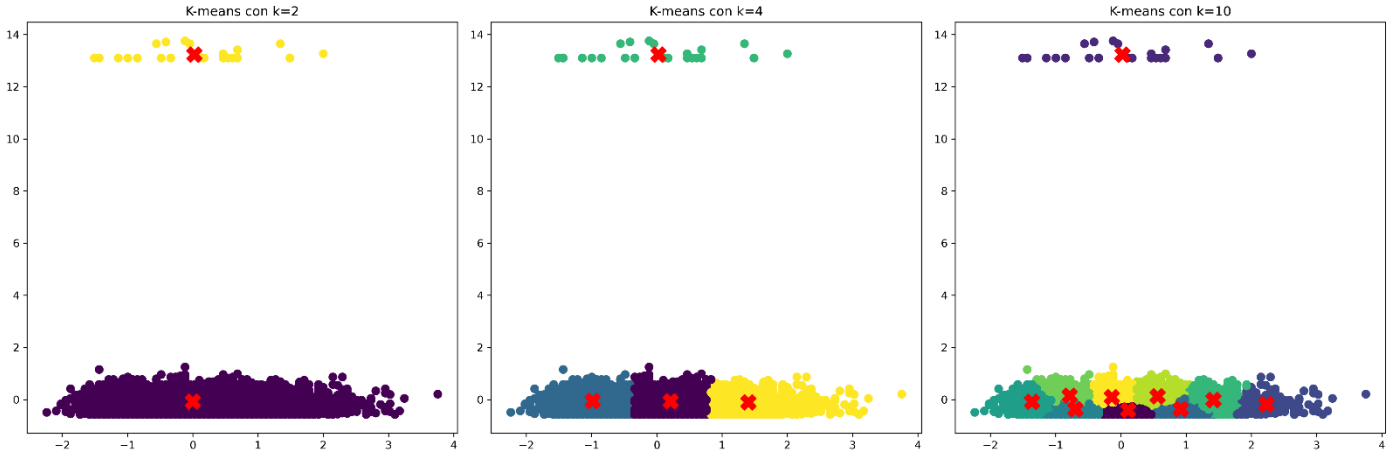
El gráfico de loadings muestra que el primer componente está fuertemente influido por la edad (CH06 y edad2), mientras que el segundo componente está más relacionado con salario semanal, educación y horas trabajadas. Las direcciones similares de CH06 y edad2 confirman su alta correlación. En cambio, salario y educación aportan más varianza en un eje distinto. Esto sugiere que los componentes capturan dimensiones distintas: una demográfica y otra laboral/educativa.

4)



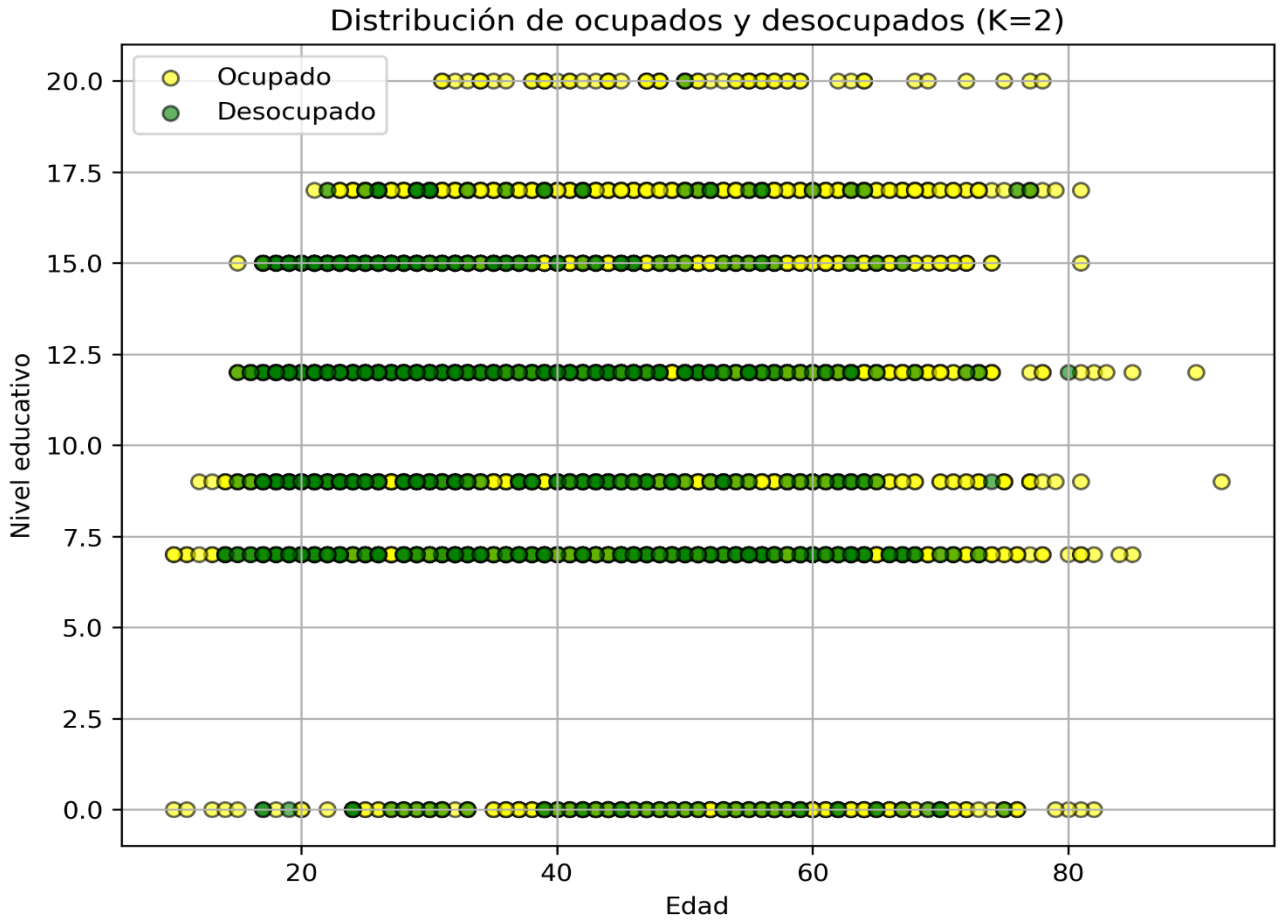
Este gráfico muestra la proporción de varianza explicada por cada componente principal en un análisis de PCA. Se observa que los primeros tres componentes explican la mayor parte de la varianza del conjunto de datos, con una acumulación cercana al 85%. A partir del cuarto componente, la ganancia de varianza explicada es mínima, lo cual sugiere que podrían descartarse sin perder mucha información. Esto indica que una reducción a tres componentes sería razonable para simplificar el análisis manteniendo la mayoría de la información.

5) a)



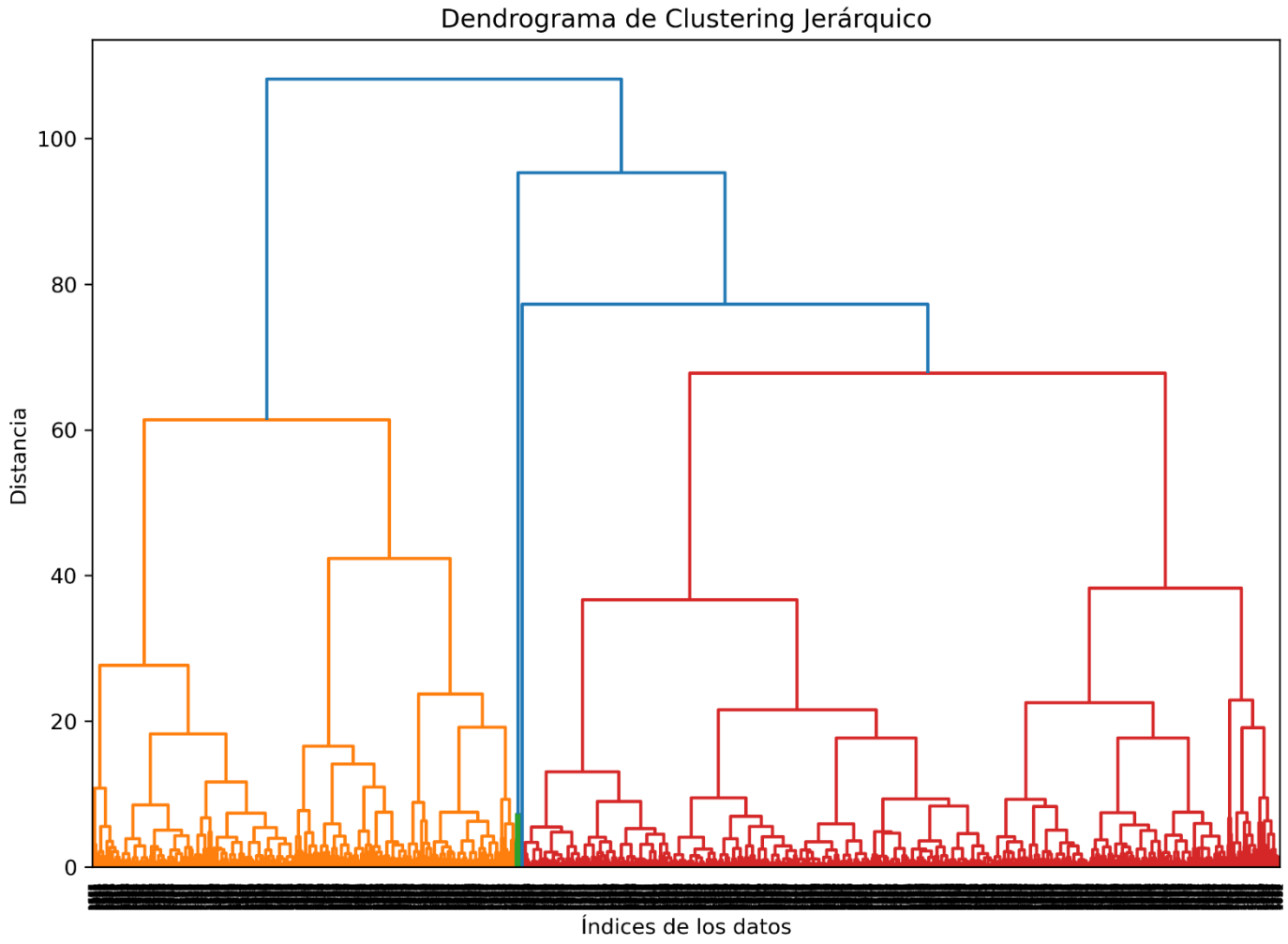
Este conjunto de gráficos muestra los resultados de aplicar K-means con diferentes valores de *k* (2, 4 y 10). Con *k=2*, los datos se dividen en dos grupos amplios, lo que puede ser insuficiente para capturar la estructura real. Al aumentar a *k=4*, los clusters se ajustan mejor a distintas zonas del espacio. Con *k=10*, los grupos son más específicos, pero puede observarse una posible sobresegmentación, lo que podría llevar a una menor interpretabilidad.

b)



El gráfico muestra la distribución de personas ocupadas y desocupadas según su edad y nivel educativo, utilizando dos grupos (K=2). Se observa que ambos grupos están presentes en todos los niveles educativos y rangos de edad, lo que sugiere que no hay una separación clara entre ocupación y estas variables. Sin embargo, podría notarse una leve concentración de ocupados (amarillo) en niveles educativos más altos. La superposición de puntos indica que otros factores podrían estar influyendo en el estado laboral.

6)



El dendrograma muestra la estructura jerárquica de los datos mediante agrupamiento, permitiendo visualizar cómo se forman los clústers al aumentar la distancia de enlace. Se observan claramente tres grandes grupos principales antes de una gran fusión, lo que sugiere que una partición en tres clústers podría ser apropiada. Además, las alturas a las que se unen los grupos indican qué tan diferentes son entre sí. Esta visualización es útil para determinar el número óptimo de clusters en un análisis jerárquico.