# Informe

## Trabajo Práctico N°3

### Big Data y Machine Learning

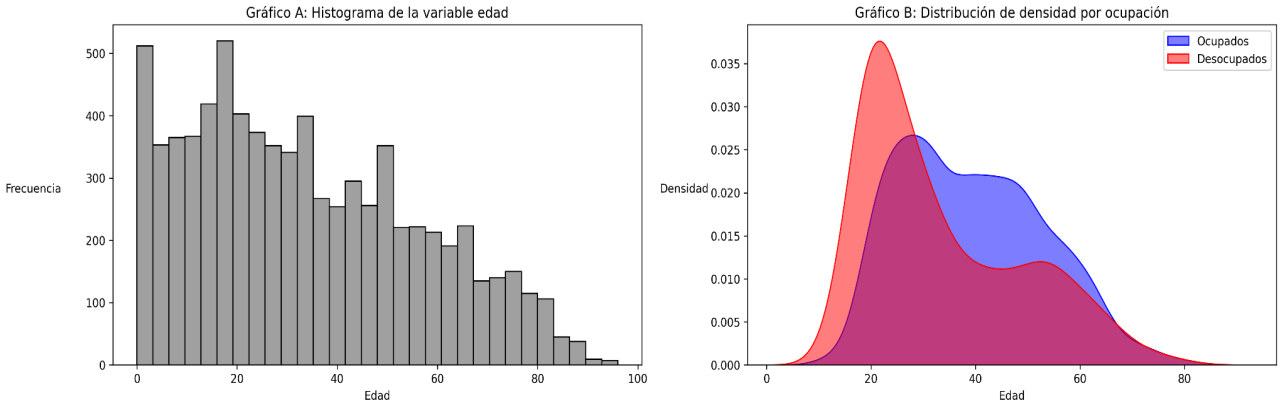
Grupo 24: Agustín Cané, Joaquín García Lucchesi,Joaquín Tesler

Facultad de Ciencias Económicas, Universidad de Buenos Aires

Mayo 2025

1) La muestra utilizada para el análisis de este trabajo proviene de datos de la EPH correspondientes al primer trimestre de 2004 y 2024 para el Gran Buenos Aires. Los siguientes gráficos muestran la distribución de la variable edad dentro del total de la muestra y la predominancia de la ocupación y la desocupación según la edad de los encuestados.

Graficos 1




Con respecto al gráfico A, se aprecia una distribución decreciente de la edad de la población, con una inflexión en torno a los 50 años. Se reconoce igualmente un pico alrededor de los 10 y 30 años. Esto nos da la pauta de que la muestra presenta una mayoría de personas en edad activa (o cerca de ella).

En el gráfico B, vemos en rojo la distribución de densidad de los desocupados y en azul la de los ocupados. La distribución es más extendida en edad para los ocupados, comprendiendo desde los 20 a los 60 años de forma sostenida.

Por el lado de los desocupados, se reconoce un pico alrededor de los 20 años, que luego decrece pronunciadamente hasta llegar a los 40. Esto sugiere que donde se presenta una mayor preponderancia de la desocupación según los registros consultados es en la población joven.

2) La tabla 1 da a conocer una estadística descriptiva de los años de educación formal a partir de los relevamientos de la EPH para el primer trimestre de 2004 y 2024 tomados en conjunto. Los datos exhibidos se corresponden con el número de años de educación según el nivel educativo, asociados con el siguiente criterio:

- Terminada la escuela primaria se cuenta con 6 años de educación formal completados.
- Para ex-alumnos de EGB, se cuentan 9 años de instrucción al terminar esta instancia . Todos los niveles inferiores de educación no son tenidos en cuenta al momento de sumar los años de escolaridad.
- Se atribuye a quienes terminaron la escuela secundaria /polimodal 12 años de educación formal. A quienes completaron su educación terciaria les corresponden 15, universitaria 17 y educación de posgrado 20.

Tabla 1: Estadística descriptiva sobre los años de educación formal

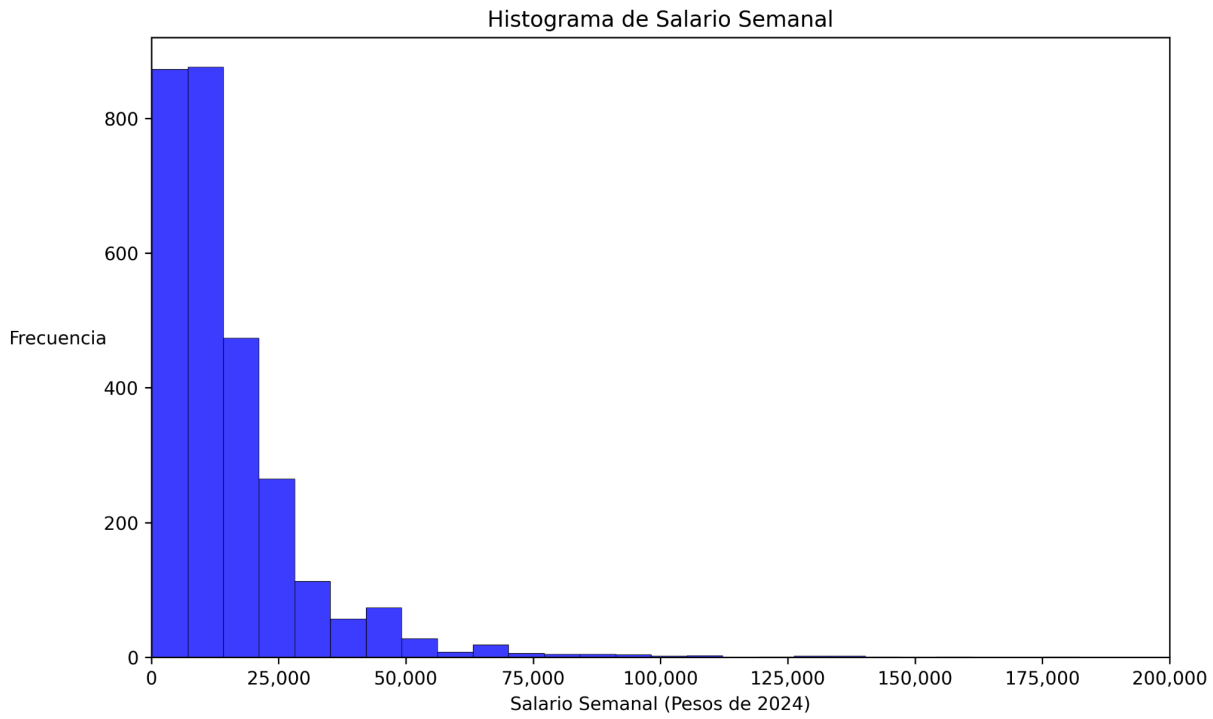
| | Total Observaciones | Media | Desvío Estándar | Mínimo | Máximo | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|---|
| Años de educación formal | 9.360 | 10,1 | 4,0 | 0 | 21 | 6 | 12 | 13 |

En base a estos datos , se puede concluir que :

- 21 años es el máximo tiempo de educación formal declarado por una persona dentro de la muestra
- El percentil 25 de la muestra solo ha terminado como máximo los estudios primarios
- El percentil 50 de la muestra ha terminado como máximo los estudios secundarios
- El percentil 75 de la muestra ha terminado sus estudios secundarios y no ha concluído su educación universitaria/terciaria

3) El siguiente gráfico describe la distribución del salario semanal expresado en pesos de 2024 para la muestra de nuestro análisis.

Grafico 3

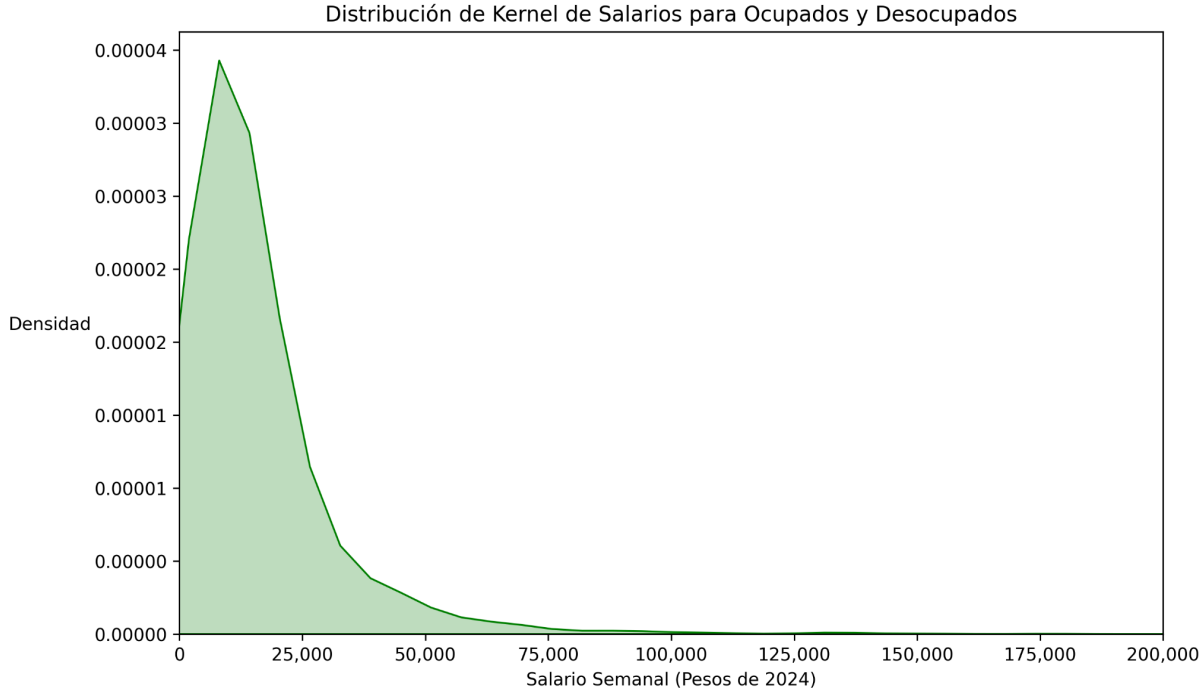


A fin de preservar la claridad expositiva, tanto en el gráfico 3 como en el gráfico 4 se representa la información poniendo como máximo en la escala del eje de abscisas un salario de $ 200.000 semanales.
En el eje X vemos el valor del salario semanal, ajustado a pesos de 2024. En el eje Y, la frecuencia, es decir la cantidad de observaciones en las que se repite cierto valor. Vemos que estas se concentran en los valores que van entre 0 a $25.000.  Se reconocen igualmente bins de mayor extensión a la izquierda del punto correspondiente a los $25.000 semanales, lo que indica que existe dentro de la muestra una parte mayoritaria de los encuestados que accede semanalmente a un salario inferior a este monto.

Los valores de salarios semanales superiores a los $75.000 representan una proporción minoritaria de la muestra dado que, tal como muestra el gráfico 3,  su frecuencia es mínima.

El gráfico 4 ilustra, para la misma base de datos, cómo se distribuye en proporción el salario entre ocupados y desocupados.

Grafico 4




En el eje de abscisas vemos nuevamente representado el valor de salario semanal ajustado a pesos de 2024. El  eje de ordenadas representa la densidad.

A pesar de esta diferencia, el resultado es muy similar al histograma : Vemos un pico positivo entre los valores de 0 a $25.000, que decae rápidamente. El pico máximo se aproxima más al 0 que a $25.000, indicando una mayor densidad de observaciones inferiores a este monto.

La caída rápida de la densidad en la distribución de Kernel coincide con la caída de frecuencia del histograma , en torno a los $75.000 semanales.

4) La tabla 2 muestra una estadística descriptiva de las horas de trabajo total de los sujetos encuestados en la muestra de nuestra muestra analizada.

Los datos se muestran acorde a la duración de una semana hábil , no pudiendo exceder los registros de la muestra las 120 horas trabajadas por semana para ser tenidos en cuenta.

Tabla 2 : Estadística descriptiva sobre las horas trabajadas semanalmente

| | Total observaciones | Media | Desvío Estándar | Mínimo | Máximo | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|---|
| Horas Trabajadas | 14.643 | 15,43 | 22,66 | 0 | 114 | 0 | 0 | 35 |

En base a esta información, podemos concluir que:

- El máximo de horas trabajado declarado en los relevamientos es de 114 horas por semana
- En promedio los encuestados trabajan aproximadamente 15 horas a la semana
- El percentil 75 de los encuestados trabaja como máximo 35 horas por semana
- La mitad de los encuestados declaró no trabajar

5)

Tabla 3 : Resumen de la base de datos final para la Región 1, Gran Buenos Aires

| | 2004 | 2024 | Total |
|---|---|---|---|
| Cantidad de Observaciones | 7.647 | 7.647 | 14.698 |
| Cantidad de NAs en la variable “Estado” | 0 | 0 | 0 |
| Cantidad de Ocupados | 3.079 | 3.224 | 6.303 |
| Cantidad de Desocupados | 528 | 311 | 839 |
| Cantidad de variables limpias y homogeneizadas | 31 | 31 | 31 |

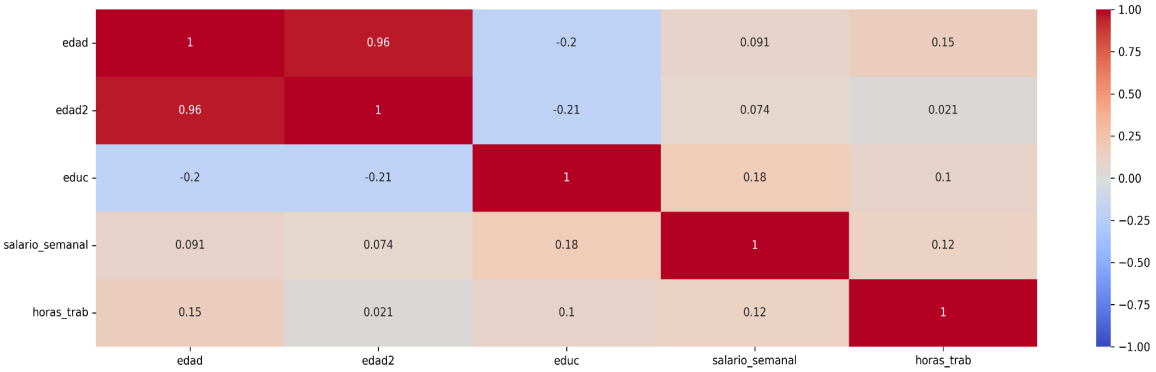
Para ambos años corresponde el mismo número de variables homogeneizadas , las cuales conforman el total de variables registradas en la base de datos final. Igualmente, el número de observaciones es el mismo en los dos registros.

Las principales diferencias se aprecian en la cantidad de desocupados y ocupados, correspondiéndole una menor proporción de los primeros y una mayor de los segundos a la EPH de 2024. Por otra parte, los datos analizados muestran que en ambos relevamientos todos los encuestados respondieron acerca de su condición laboral, lo que explica que no se hayan obtenido valores NA (vacíos) para la variable “Estado”.

## Parte 2

1) El siguiente heatmap describe la correlación entre las variables que representan la edad, el número de años de educación formal, el salario semanal y las horas trabajadas de los individuos encuestados.

Grafico 5

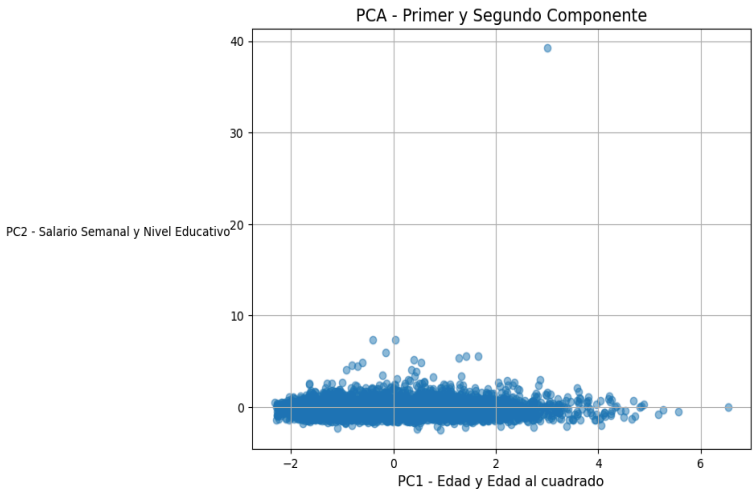



Se reconoce, a partir de los tonos más cálidos en los recuadros, un vínculo más estrecho entre las horas trabajadas y el nivel educativo, en contraste con el resto de registros. Le sigue en escala el salario semanal,mostrando igualmente que existe una dependencia entre los valores que toman ambas variables.

Por otra parte, los recuadros de color más oscuro, y por ende mayor correlación, entre variables distintas entre sí corresponden a la edad y “edad2”. Esto se explica a partir de que ambas toman los mismos datos para formar sus resultados, aunque elevándolos a exponentes distintos.

2) El Gráfico 6 ilustra los resultados de un análisis de principales componentes aplicado sobre las variables mencionadas en el punto anterior.

Grafico 6



El gráfico de dispersión relaciona las observaciones existentes de los componentes principales. En el eje de las abscisas encuentran aquellas observaciones del primer componente, mientras que en el de las ordenadas, las del segundo componente. Tras un análisis de los ponderadores (loadings), se obtuvieron los siguientes datos :

Tabla 4 : Ponderadores según variable y componente principal

| Variables | PC1 | PC2 |
|---|---|---|
| Años al momento de la encuesta | 0,699835 | 0,038868 |
| Edad (al cuadrado) | 0,699915 | 0,023418 |
| Educación | -0,106625 | 0,530400 |
| Salario Semanal | 0,048495 | 0,787732 |
| Horas Trabajadas | -0,081417 | 0,309993 |

En base a esta información, se puede concluir que en el PC1 tienen mayor incidencia la Edad y la Edad al Cuadrado, mientras que en el PC2, el Salario Semanal y el Nivel Educativo.

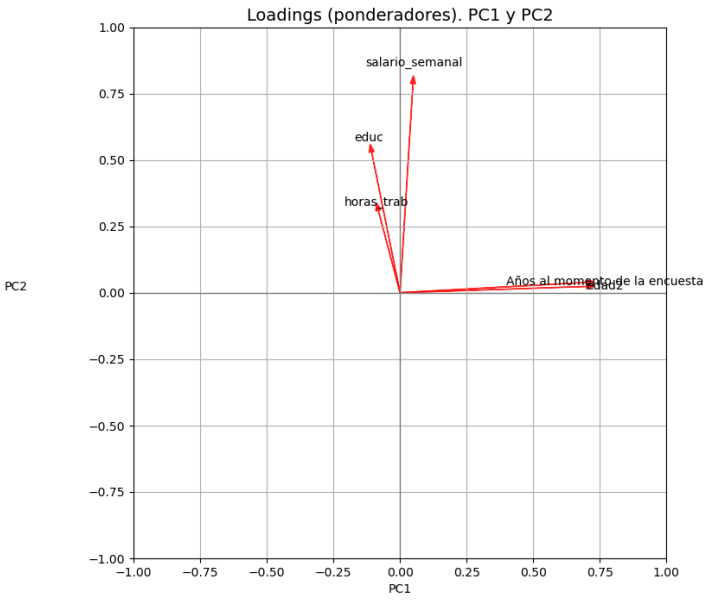
Si los puntos se agrupan en la parte superior derecha del gráfico, tienen valores altos en ambos componentes, lo cual indicaría que las observaciones comparten atributos que las colocan en los lugares de mayor varianza de ambos componentes.

Lo opuesto se podría decir si se agrupan en la parte inferior izquierda, como sucede en este caso; tenemos un conjunto denso de observaciones que se agrupa en los valores bajos de varianza de ambos componentes.

La información volcada en el gráfico permite establecer que, dada la agrupación de puntos en la parte inferior izquierda del esquema, las observaciones comparten características similares y coherentes con una menor cantidad de años al momento de la encuesta, un menor salario semanal o un menor nivel educativo.

3) El gráfico 7 da a conocer mediante un diagrama de flechas el vínculo entre variables y ponderadores dentro del modelo.

Grafico 7

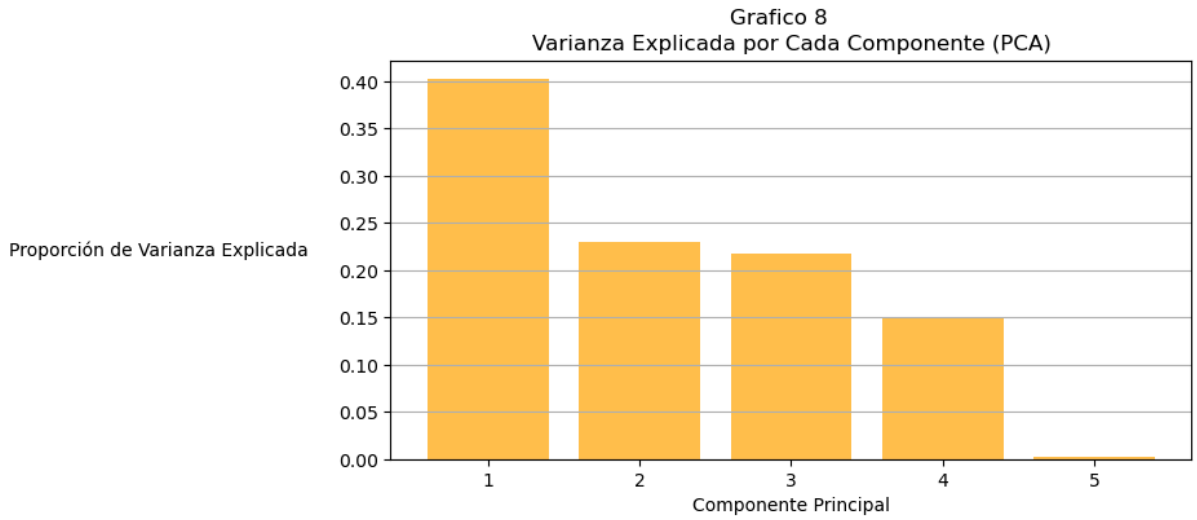



En el eje de las abscisas se expresa el valor de los ponderadores para el primer componente principal, mientras que en el eje de las ordenadas, el valor de los ponderadores para el segundo componente principal.

La dirección de las flechas indica su correlación positiva o negativa con el componente. Como vemos, las horas trabajadas, el salario semanal y la educación tienen una correlación positiva con el segundo componente, mientras que la edad y la edad al cuadrado tiene una correlación positiva con el primer componente.

Esto nos permite afirmar nuestra conclusión anterior. Por otro lado, el tamaño de la flecha indica la magnitud del loading (ponderador), siendo mayor la magnitud en el caso de la edad, pero menor en el caso del resto de variables.

4) El gráfico 8 ilustra el grado de injerencia que cada uno de los elementos del PCA ejerce sobre la varianza




El eje de abscisas muestra los componentes principales. El eje de ordenadas, la proporción de varianza explicada por cada componente.

Con esta información se reconoce que un 40% de la varianza total de los datos está explicada por el primer componente, a saber la edad y la edad al cuadrado.

Por otra parte, existe una muy leve diferencia entre la proporción correspondiente al segundo y tercer componente. Esto nos permite concluir que hay más dimensionalidad en el análisis del PCA, es decir, los componentes principales no son suficientes para explicar toda la varianza. Es más, vemos que el cuarto componente también explica una porción significativa de la varianza.

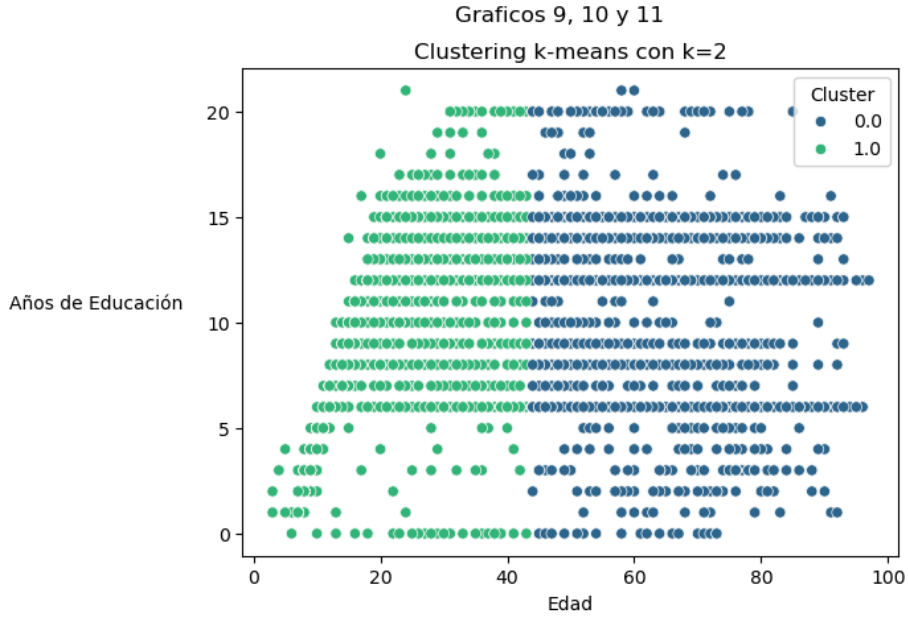
Puede entonces que el análisis de la varianza entonces adquiere cierta complejidad, ya que es necesario sumar al análisis más dimensiones, en este caso los PCA n°3, 4 y 5, para explicar el total de la varianza. Se visualizan los datos de los loadings en la siguiente tabla:

Tabla 5: Ponderadores según variable y componente principal

| Variables | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Años al momento de la encuesta | 0,699835 | 0,038868 | -0,003193 | -0,092846 | 0,707170 |
| Edad (al cuadrado) | 0,699915 | 0,023418 | -0,006615 | -0,098786 | -0,706942 |
| Educación | -0,106625 | 0,530400 | -0,613721 | -0,575022 | -0,001901 |
| Salario Semanal | 0,048495 | 0,787732 | 0,105701 | 0,604834 | -0,011400 |
| Horas Trabajadas | -0,081417 | **0,309993** | **0,782381** | -O,533991 | -0,003041 |

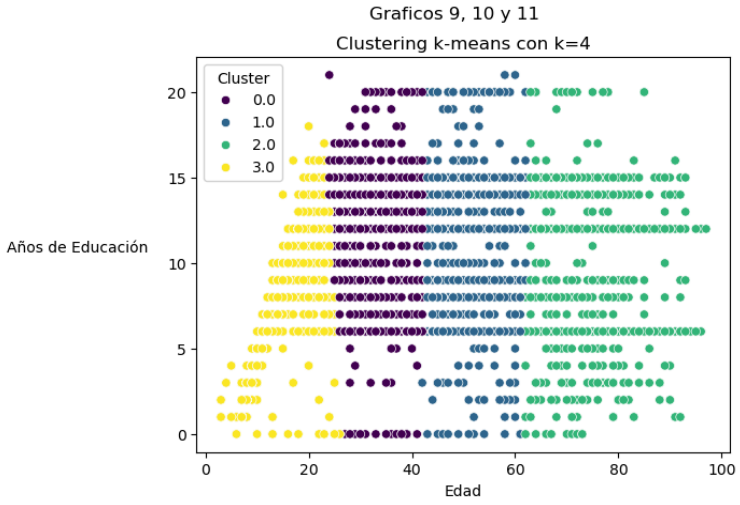
Las horas trabajadas adquieren un peso importante, sobre todo en el tercer componente. Estas entonces explican prácticamente la misma proporción de varianza que el segundo componente, es decir el salario semanal y el nivel educativo.

5) Los siguientes gráficos muestran un análisis de agrupación en clusters con distintos niveles de desagregación hecho a partir de las variables identificadas en el ejercicio anterior.



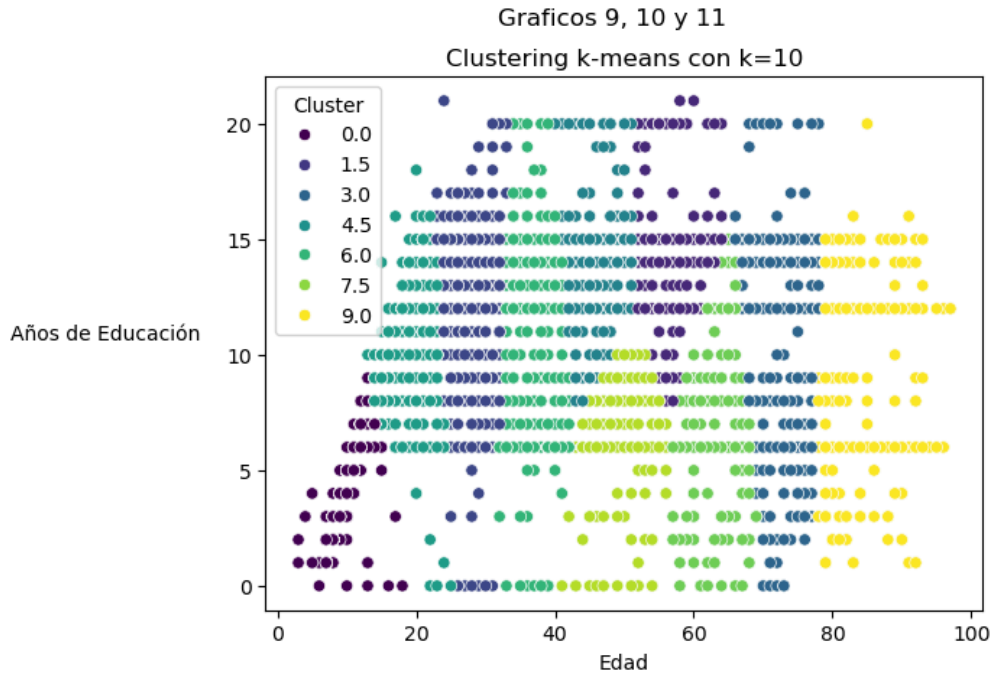

El cluster con k = 2 separa en dos grandes grupos la información. En el eje de las X vemos la edad de las observaciones, en el eje de las Y, los años de educación formal. Hay una separación marcada, entre los que van del rango etario de 0-40 años, y el resto que van de 40-100 años (mayor edad). En promedio podemos intuir, en base a la condensación de datos, que el algoritmo separa en un grupo de jóvenes entre 15-40 años que tienden a tener más años de educación (cluster verde) que el grupo que va de 50-80 años (cluster azul).

Ahora, veamos que sucede cuando k=4;



El cluster con k = 4 separa en cuatro grupos. La segmentación es más detallada; de los 0-20 años vemos un grupo de observaciones con bajo nivel educativo (0-5 años), pero que es mucho menor al grupo con un nivel educativo superior (5-15 años). El rango de educación con mayor cantidad de datos sigue siendo el de 0 a 15 años de educación formal. Sin embargo, vemos que a mayor edad de las observaciones es menor la cantidad en ese rango, y mayor la cantidad que tiene un bajo nivel educativo (observando los cluster de color verde).

Por último, analizamos el cluster con k=10

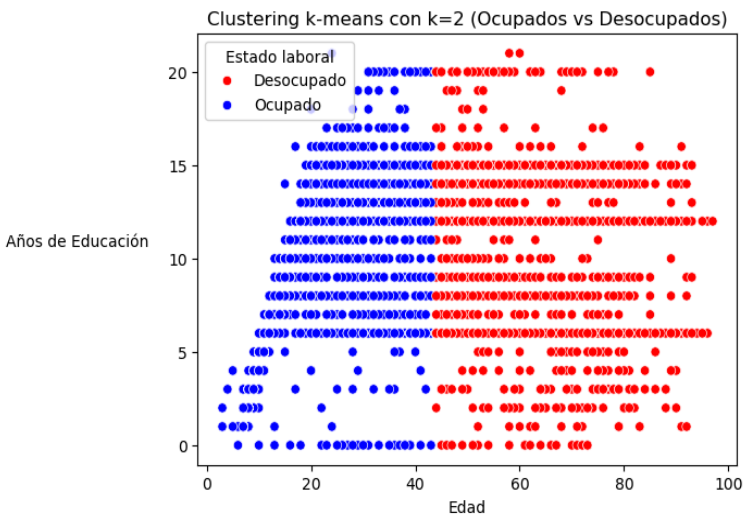



El cluster con k = 10 nos permite identificar varios subgrupos. Por ejemplo, observamos personas jóvenes con bajo nivel de educación (0 a 3 años) pero también personas mayores con bajo nivel educativo (cluster de color amarillo). Las personas adultas en edad productiva son las que presentan, en general, mayor nivel educativo. Estos son los adultos de 20 a 60 años.

6) Cluster con k=2 - ocupados y desocupados

El gráfico 12 muestra una ilustración de predicciones hechas por el algoritmo de cluster k-means con k=2, ya que está tomando en cuenta el estado laboral de dos grupos; ocupados y desocupados

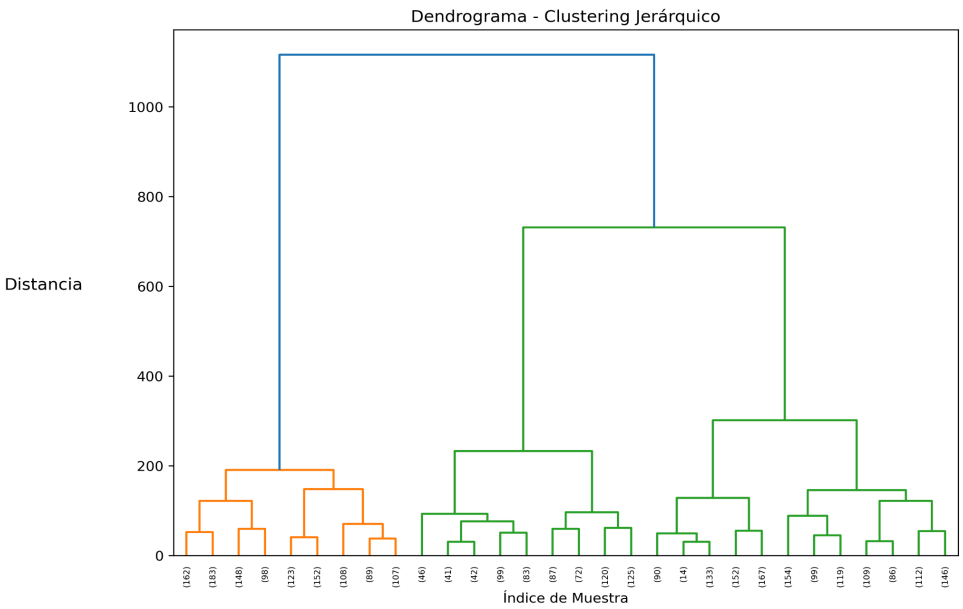
Gráfico 12




Hay una división muy marcada similar a la del gráfico de k=2. Pero ahora estamos analizando casos de ocupados y desocupados; un grupo de color azul predice los ocupados, el otro de color rojo los desocupados (esto según k-means). Hay ciertas coincidencias; podemos ver que un grupo en rojo presenta un menor nivel educativo, y estas observaciones son mayores que las del color azul. Además vemos más dispersión en los niveles educativos para las personas de mayor edad, lo cual puede ser una señal de una desocupación que es sostenida en el tiempo (ej. personas de más de 45 años con bajo nivel educativo son mucho más propensas a estar desempleadas ya que no poseen atributos productivos).

7) El gráfico 13 representa una esquematización de la división en clusters de la información analizada, cubriendo desde los agrupamientos más grandes hasta los de menor dimensión.

Grafico 13



El dendrograma representa de forma gráfica el cluster de tipo jerárquico. En el eje de las abscisas vemos los índices de las muestras: en nuestro caso había más de 14000 observaciones, por ende, seleccionamos las 30 principales para preservar la claridad expositiva. En el eje de las ordenadas se da una referencia para la distancia entre clusters. Los clusters en el dendrograma se van agrupando y forman las llamadas “ramas”, que son las uniones visibles.

Si el punto de unión está en una altura baja con respecto al eje Y (ejemplo los cuadrados pequeños en la base del gráfico) los elementos son más similares entre sí. Ante una mayor distancia, disminuye su similitud. Esto se expresa en la longitud de las “ramas”. Un ejemplo claro es la unión de color azul; sus elementos presentan mucha distancia, por ende son muy diferentes entre sí. Una función importante de estos dendrogramas es poder visualizar cuales son las relaciones principales entre muestras, a fin de simplificar el análisis de clusters con k-means.