TRABAJO PRACTICO 3-\_BIG DATA\_GRUPO 16

**Grupo** 16: Compuesto por: Maximiliano Petrini, Augusto Fasciani, Lautaro Solís.

**Motivo del trabajo de investigación:** El siguiente trabajo se realizó gracias a las encuestas realizadas, las EPH (Encuesta Permanente de Hogares), por el INDEC (Instituto Nacional de Estadísticas y Censos) en los años 2004 y 2024

Empezamos el trabajo filtrando y ordenado las bases de datos para posteriormente tratar sobre las distintas variables y sus cambios entre el 2004 y el 2024

Las variables con las que tuvimos que trabajar fueron: Edad, Edad2, Educación, Salario Semanal Y Horas Trabajadas.

***Edad:***

Son las edades de los encuestados

***Edad2:***

Serían las edades de los encuestados elevado al cuadrado.

***Educación (Edcu):***

Esta variable transforma el grado o progreso de nivel de educación del encuestado en cantidad de años educativos***.***

***Salario Semanal:***

Es lo que gana la persona por semana, para ello utilizamos la data de salario mensual y la dividimos según las semanas trabajadas.

***Horas Trabajadas:***

Utilizamos también el salario mensual y la dividimos según la cantidad de horas trabajadas que lleva cada día/o semana.

**Desarrollo del trabajo:**

Cumpliendo las consignas preestablecidas, empezamos utilizando la base de Del anterior trabajo, es decir el “TP2 DE BIG DATA-ROMERO NOELIA”.

Previamente antes de crear las nuevas variables, debimos limpiar aquellas columnas o base de datos donde aparecen varios valores sin sentido, como nulos, Nan, casillas vacías, etc. Para poder trabajar mejor los datos y que los modelos no se vean afectado. Esto ayuda a tener una estimación más limpia y precisa en los análisis que siguen.

Una de las primeras variables a analizar fue la Edad y Edad2(Edad al cuadrado). Podemos observar en el gráfico logarítmico (o panel A) muestra una gran concentración de personas jóvenes (hasta 30 años), lo que probablemente se debe a que estás trabajando con un dataset donde la mayoría son jóvenes.

El gráfico B muestra que los desocupados tienen una mayor dispersión en su distribución de edades, lo que indica que hay desocupados tanto en los rangos jóvenes como en edades mayores, mientras que los ocupados se concentran más en las edades jóvenes.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

El siguiente paso que ralizamos fue crear la variable Educación(representando como la cantidad de añosde educación). A esta variable tuvimos que presentarle una estadística descriptiva (incluyendo el promedio, desviación estándar, mínimo, median y máximo.

Estadísticas descriptivas de años de educación:

Promedio: 8.52 años

Desviación estándar: 5.04

Mínimo: 0 años

Mediana (p50): 8 años

Máximo: 21 años

A partir de esta construcción, se observó que el promedio de años de educación es de 8,27 años, con una desviación estándar de 4,87. El valor mínimo es 0 años, correspondiente a personas sin educación formal, mientras que el máximo es 21 años, lo que indicaría estudios de posgrado completos. La mediana es de 8 años, lo cual indica que al menos la mitad de los individuos en la muestra no completaron el nivel secundario.

Nuestra próxima parada en el camino de este trabajo de investigación, fue la variable de salario semanal. Esta variable nos sorprendió bastante. En primer lugar debimos transformar los pesos de 2004 a los de 2024, ya que los pesos de 2004 tienen un poder de compra distinto a los pesos de 2024 primer trimestre.

En conjunto, los gráficos sugieren una estructura salarial desigual, con una parte significativa de la población concentrada en niveles bajos y medios de ingreso, y una minoría con salarios más elevados. La bimodalidad podría reflejar diferencias por tipo de ocupación, nivel de formalidad o calificación laboral.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Por último creamos nuestra última nueva variable llamada “Horas Trabajadas”, que representa el total de horas trabajadas como la suma de las horas en la ocupación principal y otras ocupaciones.

**Comparación de estadísticas descriptivas:**

Estadísticas con valores > 168 horas:

Número de observaciones: 14

Porcentaje de datos eliminados por > 168h: 0.13%

Estadísticas finales (0 < horas ≤ 168):

Promedio: 39.60 horas

Desviación estándar: 18.46

Mínimo: 1 horas

Mediana (p50): 40 horas

Máximo: 126 horas

Número de observaciones: 10979

El análisis final, que excluye tanto los valores cero como aquellos superiores a 168 horas, nos proporcionó una distribución mucho más realista y confiable. Esta distribución muestra una clara concentración alrededor de las 40-45 horas semanales, lo cual es consistente con una jornada laboral estándar.

Decidimos eliminar valores superiores a 168, porque es matemáticamente imposible trabajar esas horas, sabiendo que una semana tiene 168 horas.

De manera gráficamente mostramos las observaciones:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2004 | 2024 | Total |
| Cantidad de Observaciones | 14821 | 14410 | 29231 |
| Observaciones con NaN en ESTADO | 0 | 0 | 0 |
| Cantidad de Ocupados | 5682 | 6302 | 11984 |
| Cantidad de Desocupados | 1035 | 453 | 1488 |

Cantidad de variables limpias y homogeneizadas: 5

Gráfico

El contenido generado por IA puede ser incorrecto.**PARTE 2**

Pasamos a las Segunda Parte del Trabajo, en donde tuvimos que utilizar las variables previamente hechas, y Aplicarles PCA.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Al ser muchos la cantidad de encuestados se puede ver que hay dispersión de los puntos por la gran parte del Gráfico. Aunque hay una mayor concentración por el centro del gráfic, es decir de -2,5 a 3 en el CP1 y de -2 a 4 en el CP2.

A su vez podemos observar que el peso de los ponderadores esta distribuido de manera distinta, ya que se puede notar como las variables Edades y Edades 2 tienen más peso que las otras variables

Loadings:

Edad: 0.66566514

Edad2: 0.62698772

Educ: 0.319649

Salario Semanal: 0.17613358

Horas Trabajadas: 0.1748651

Luego podemos ver en el siguiente gráfico la proporción de varianza explicada.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Luego pasamos a las **Parte 2**, B del trabajo que se trata de la utilización de Cluster y K medias.

Donde debimos correr el algoritmo con K=2, K=4 y K=10 usando un n\_init =

20 y graficar los resultados usando dos predictores.

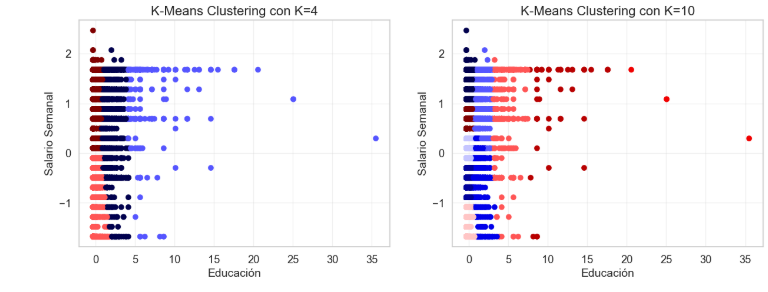


Entonces con un k = 2 se ve una clara división entre bajos ingresos y bajo salario semanal, con salario promedio o mayor con otros niveles de educación. Preferimos usar valores estandarizados para el salario a fin de que no se creen valores que generen distorsión en la interpretación, por lo que podemos entender que:

* si z = 0 -> salario promedio
* si z > 1 -> salario mayor al promedio
* si z < 1 -> salario menor al promedio

(Es decir, por ejemplo, z = 2 representa una persona que gana aproximadamente 2 desvíos estándar por encima del promedio.)

Veamos para k = 4 y k = 10



Entonces para k = 4

* Subgrupo rosa: baja o nula educación y salario entre menor y promedio.
* Subgrupo rojo: baja o nula educación y salario entre mayor y promedio.
* Subgrupo azul oscuro: educación entre 0 y 5 años, grupo con salario entre promedio y mayor.
* Subgrupo celeste: grupo con salario entre promedio y mayor, el nivel de educación esta dispersado, pero oscila entre 5 y 15 años en promedio.

Para k = 10 podemos hacer una interpretación similar, aunque la diferencia entre los distintos conjuntos se relativiza en diferencias sustanciales en el salario. Por lo tanto, con k = 4 es más fácil dar una división entre perfiles "económico-educativos" respecto a la educación y el salario semanal.

Ahora debimos graficar Edad y Educación de los resultados de 𝑘 = 2 donde cada punto tome dos colores, un color para ocupados y otro color para desocupados.

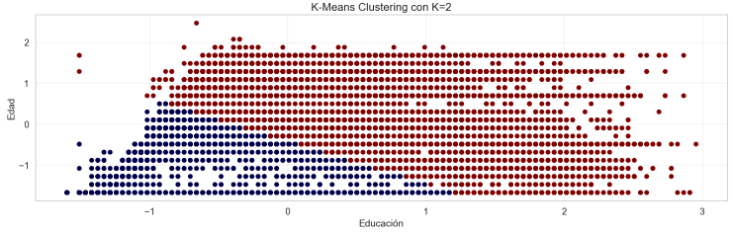


Imagen que contiene Diagrama

El contenido generado por IA puede ser incorrecto.

Podemos ver en el gráfico que hay una gran concentración de puntos rojos (ocupados) y que los puntos azules (desocupados) se encuentran dispersos y sin una agrupación clara. O sea, no hay una relación clara entre la combinación de edad y educación respecto de la ocupación. Entonces el algoritmo K-means no logra separar adecuadamente a ocupados de desocupados, porque no existe una limitación que habilite la distinción entre ambos grupos (es decir, los predictores edad y educación no pueden definir correctamente que alguien este ocupado o no).

Gráfico

El contenido generado por IA puede ser incorrecto.

El dendograma permite ver gráficamente los subgrupos para cada número de clústeres. En este caso, muestra cómo se agrupan las personas según su edad y educación. No las grafica directamente si no qué tan parecidas son entre sí. El eje X son las personas, y el eje Y muestra cuán distintas son cuando se las agrupa. Mientras más alto se juntan, más distintas son