Materia: Taller de Programación
Grupo 2: Jimena Teran, Juan Lynch, Renzo Falciglia
Fecha: 29/11/2024

Trabajo Práctico 3

Parte I: Analizando la base

Punto 1

De acuerdo a los informes técnicos que publica el INDEC sobre el mercado de trabajo e ingresos, una persona desocupada se refiere a aquella persona que, no teniendo ocupación, se encuentra buscando activamente trabajo y está disponible para trabajar. Esto se conoce como “desocupación abierta”.

Esta definición no tiene en cuenta otras formas de precariedad laboral como personas con trabajos transitorios mientras buscan otro tipo de trabajo, personas que trabajan jornadas involuntariamente por debajo de lo normal o personas que han suspendido su búsqueda laboral por falta de oportunidades visibles, entre otras. Estas otras modalidades el INDEC las calcula en indicadores separados.

La tasa de desocupación se calcula entonces como el cociente entre la población desocupada y la población económicamente activa.

Punto 2

Inciso A y B

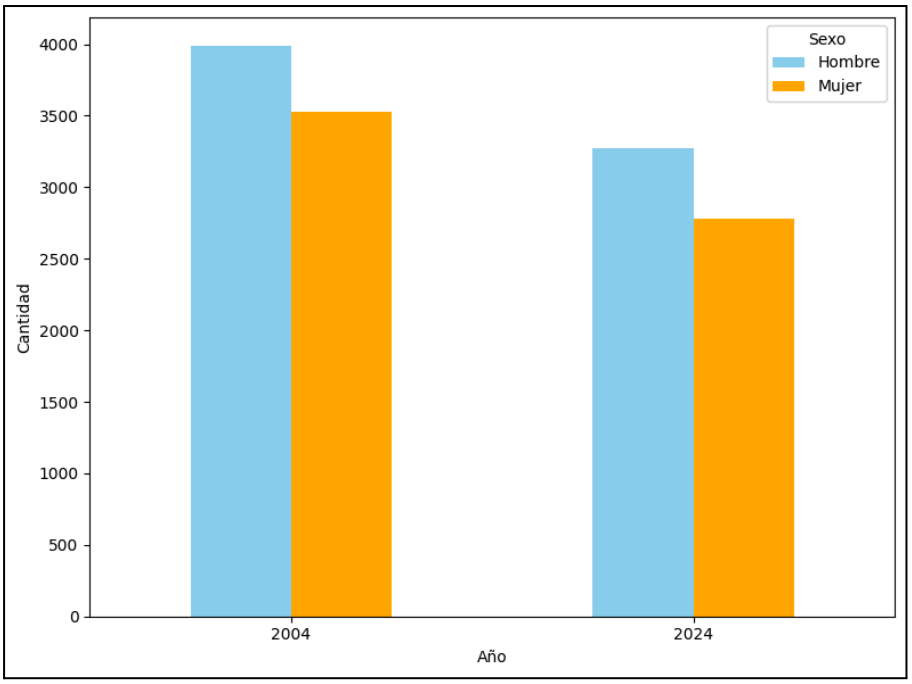
Una vez obtenido el data frame con todas las observaciones del primer trimestre de 2004 y con el primer trimestre de 2024, se eliminaron todas aquellas observaciones que no correspondan a CABA o a partidos del Gran Buenos Aires. Luego, se tomó la decisión de excluir aquellas observaciones que registraran valores negativos ya sea en la edad de las personas o en los ingresos.

A su vez, se forzó a que la columna de ingreso (IPCF) sea numérica, asignando un “NaN” en aquellos casos con valores erróneos como pueden ser caracteres de texto o faltantes. Esto último se realizó para evitar problemas con los cálculos posteriores que se realizan con dicha variable.

Inciso C

A continuación se expone un gráfico de la composición de la muestra dividida entre hombres y mujeres para cada uno de los trimestres considerados. Como se muestra en el gráfico, si bien las cantidades absolutas disminuyeron en 2024, no se observan grandes cambios en la composición. En ambas muestras se observa un mayor número de hombres.

Gráfico 1 - Composición por sexo en 2004 y 2024

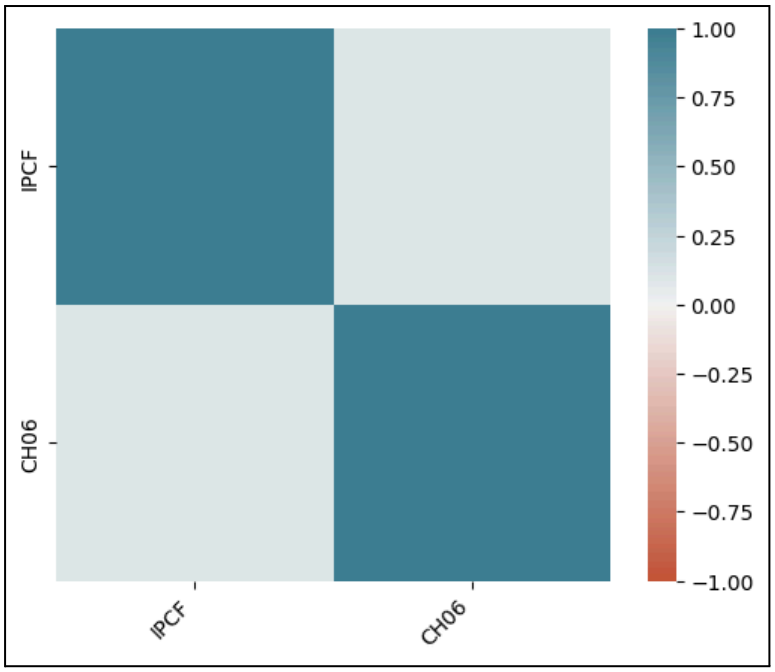


Inciso D

También se realizó una matriz de correlación entre las variables CH04, CH06, CH07, CH08, NIVEL ED, ESTADO, CAT_INAC, IPCF. Se consideró el código de la página Better Heatmaps and Correlation Matrix Plots in Python de Medium.

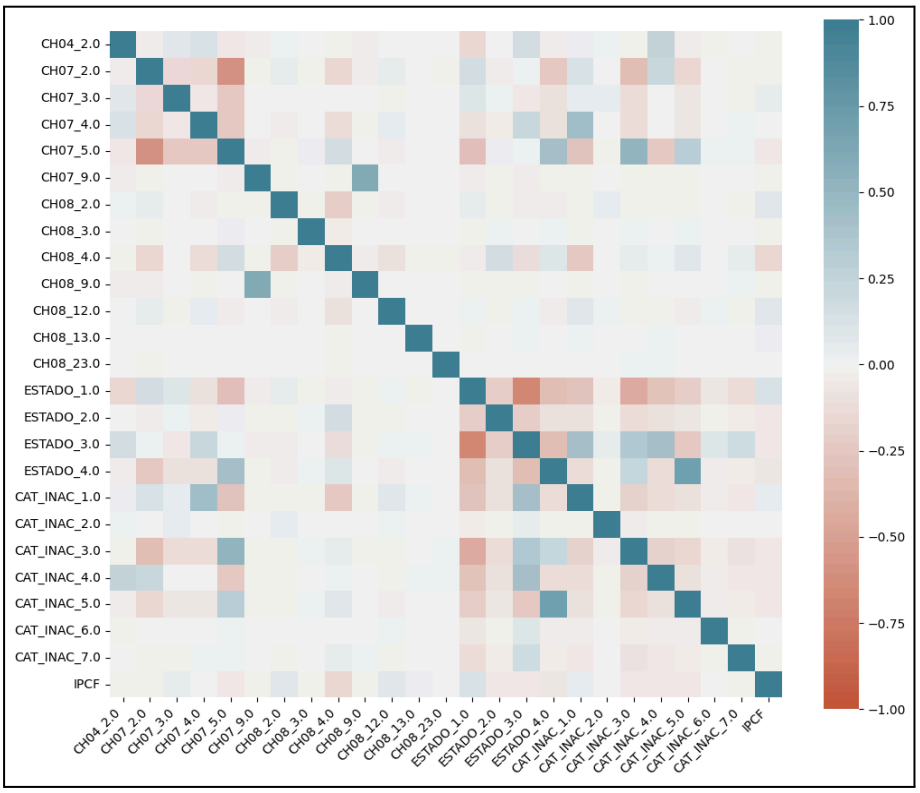
No se realizó una matriz de correlación con todas las variables mencionadas al mismo tiempo porque muchas de ellas son categóricas, sino que se realizó una primera matriz de correlación entre las variables numéricas de edad (CH06) y el ingreso per cápita familiar (IPCF), para luego realizar la técnica de One Hot Encoding con el resto de las variables categóricas.

Gráfico 2 - Matriz de correlación variables numéricas



Como se desprende de esa matriz, la edad y el IPCF no tienen una correlación fuerte. A continuación, se expone la matriz de correlación de las variables categóricas y el ingreso per cápita familiar.

Gráfico 3 - Matriz de correlación variables categóricas (one hot encoding)



Como se desprende de esta matriz, la variable de IPCF no presenta correlaciones fuertes con prácticamente ninguna variable. La correlación más fuerte que se observa es con la variable CH08, que se refiere al tipo de cobertura médica que tiene la persona. En aquellos casos donde la persona no paga ni le descuentan del sueldo ningún concepto de cobertura médica, esto se correlaciona de manera negativa por el IPCF. Esta situación implica que las personas que no abonan ningún concepto por salud tienen un ingreso per cápita familiar menor que el resto. Esto puede estar asociado probablemente a hogares pobres.

También se observa una leve correlación positiva entre el IPCF y aquellas personas que están ocupadas, lo cuál es el resultado esperable.

Inciso E

El número total de desocupados de la muestra que considera ambos trimestres es de 839, mientras que la cantidad de inactivos es de 5.462. En cuanto al ingreso familiar por categoría, a continuación se expone para cada año.

Tabla 1 - IPCF por estado

Estado	IPCF I.04	IPCF I.24
Ocupado	476	284.879
Desocupado	224	76.042
Inactivo	315	125.123

Los resultados que expone la tabla son esperables, ya que las personas con un mayor ingreso per cápita familiar son las que se encuentran ocupadas, mientras que aquellas personas con menor ingreso per cápita familiar son las que se encuentran desocupadas.

Punto 3

En este punto se calculó la cantidad de personas que no respondieron la encuesta para cada uno de los trimestres, con estos resultados:

- Cantidad de personas que no respondieron su condición de actividad en 2024: 41
- Cantidad de personas que no respondieron su condición de actividad en 2004: 10

A su vez, se dividió el data frame entre aquellas personas que respondieron y las que no respondieron.

Punto 4

En este punto se agregó una columna PEA a la base, asignando el valor 1 a quienes están ocupados o desocupados según la columna ESTADO.

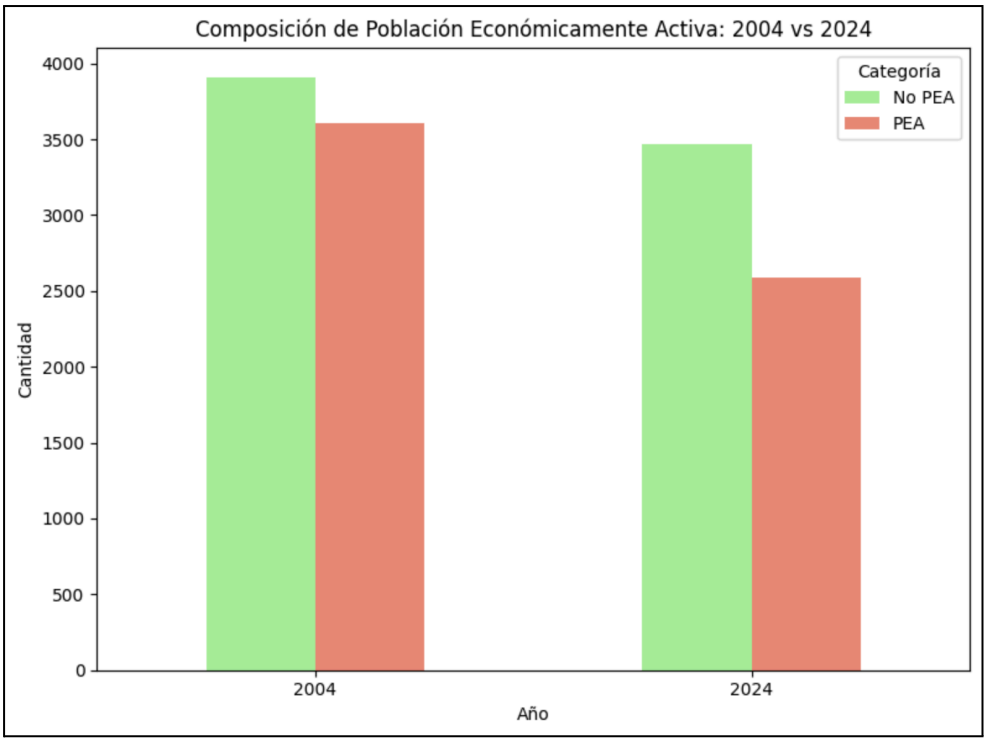
Los datos se filtraron para 2004 y 2024, identificando las personas dentro y fuera de la PEA.

Se creó un gráfico de barras mostrando la composición de la PEA en ambos años.

Resultados: En 2004, la **Población Económicamente Activa (PEA)** era considerablemente mayor a la de 2004 y mantenía un equilibrio más cercano con la población inactiva. Sin embargo, en 2024 se observa una reducción significativa en la PEA acompañada de un aumento en la población no activa.

Este cambio podría estar relacionado con factores demográficos, como el envejecimiento poblacional, o socioeconómicos, como la falta de oportunidades laborales. Además, el incremento de la población inactiva podría indicar barreras para acceder al mercado laboral, así como un mayor número de estudiantes o personas jubiladas.

Estos resultados destacan la importancia de analizar cómo variables como el nivel educativo, el género y las políticas económicas han influido en estas tendencias, para comprender mejor las dinámicas laborales y proponer estrategias que incentiven una mayor participación en el mercado laboral.



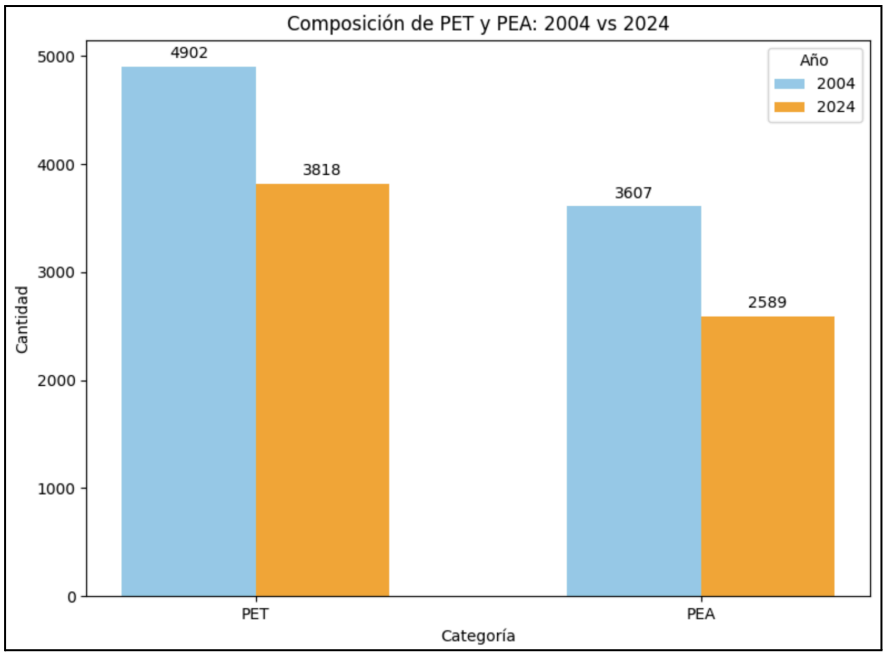
Punto 5

En el Punto 5, se añadió una columna llamada PET (Población en Edad para Trabajar), que identifica a las personas con edades entre 15 y 65 años. Se segmentaron los datos por años para analizar la composición de la PET en 2004 y 2024, y se calculó la cantidad de personas dentro y fuera de esta categoría. Posteriormente, se visualizó la distribución mediante un gráfico de barras comparativo.

En 2004, la PET era de 4902 personas, mientras que en 2024 disminuyó a 3818, reflejando un cambio demográfico significativo. La PEA (Población Económicamente Activa) también

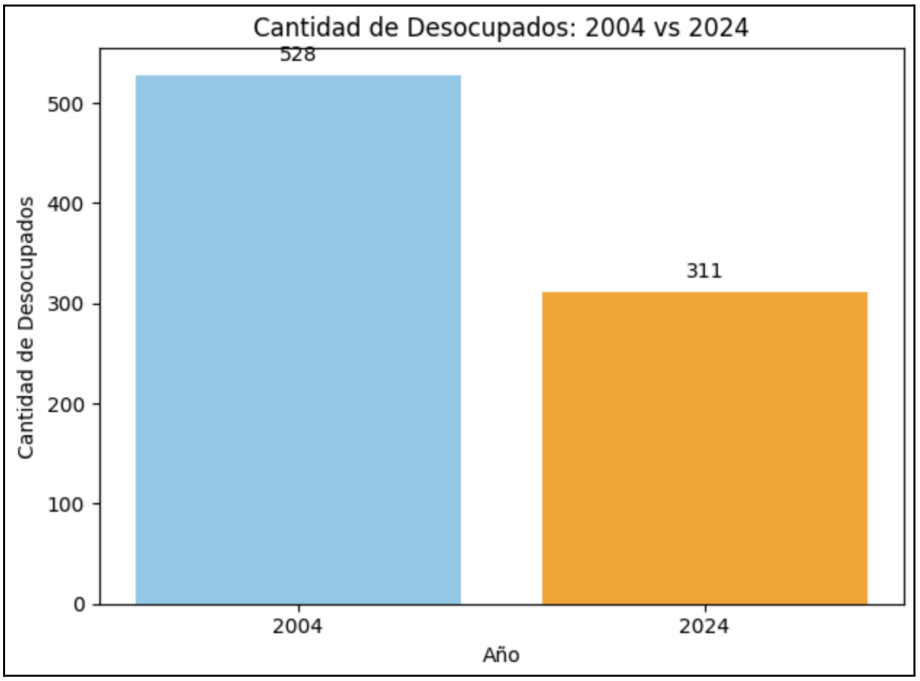
muestra una reducción, pasando de 3607 personas en 2004 a 2589 en 2024. Esta caída más pronunciada en la PEA respecto a la PET indica una menor participación laboral dentro de la población en edad de trabajar.

En 2004, el 73.6% de la PET participaba en la PEA, mientras que en 2024 esta proporción cayó al 67.8%. Estos resultados sugieren un aumento de la inactividad laboral entre las personas en edad de trabajar, posiblemente debido a jubilaciones, estudios o desincentivos laborales. El descenso en la PEA podría estar vinculado a dinámicas económicas y sociales, como la falta de empleo o condiciones laborales desfavorables. Además, la reducción de la PET puede deberse al envejecimiento poblacional, que disminuye la cantidad de personas en edad de trabajar. En conjunto, la relación entre PET y PEA refleja un cambio estructural en la composición de la fuerza laboral entre 2004 y 2024.



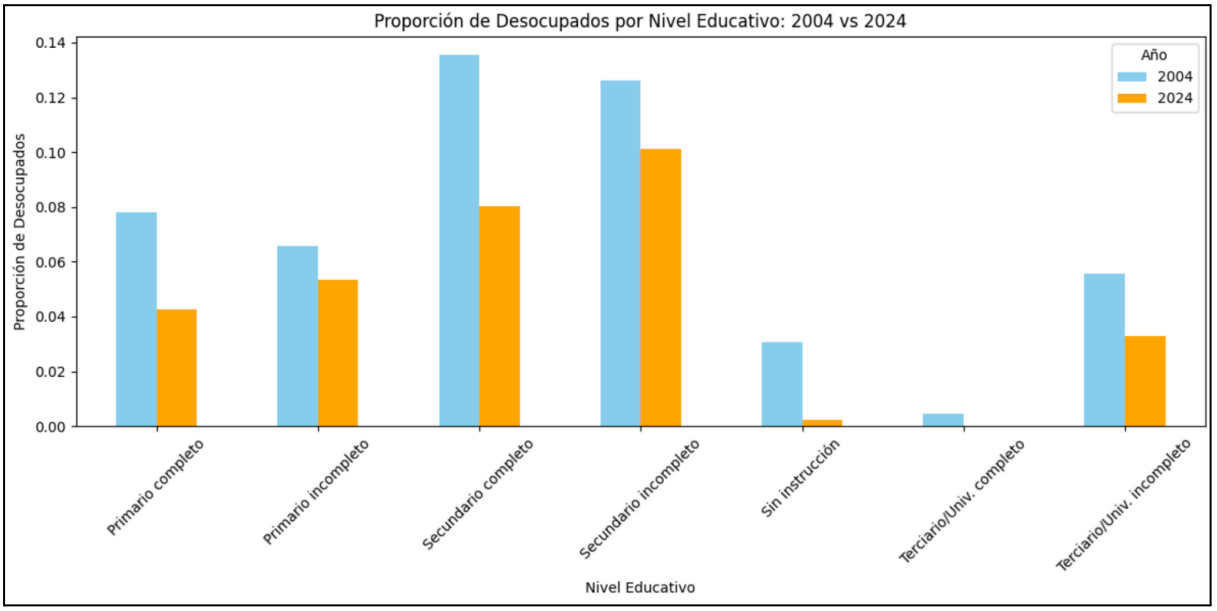
Punto 6

En este punto se creó la columna `desocupado`, asignando 1 a las personas desocupadas y `0` en caso contrario. Con esta columna, se calculó el número total de personas desocupadas en 2004 y 2024, mostrando una disminución significativa en 2024, reflejando una mejora en el mercado laboral.



Punto 6a (Desocupados por nivel educativo):

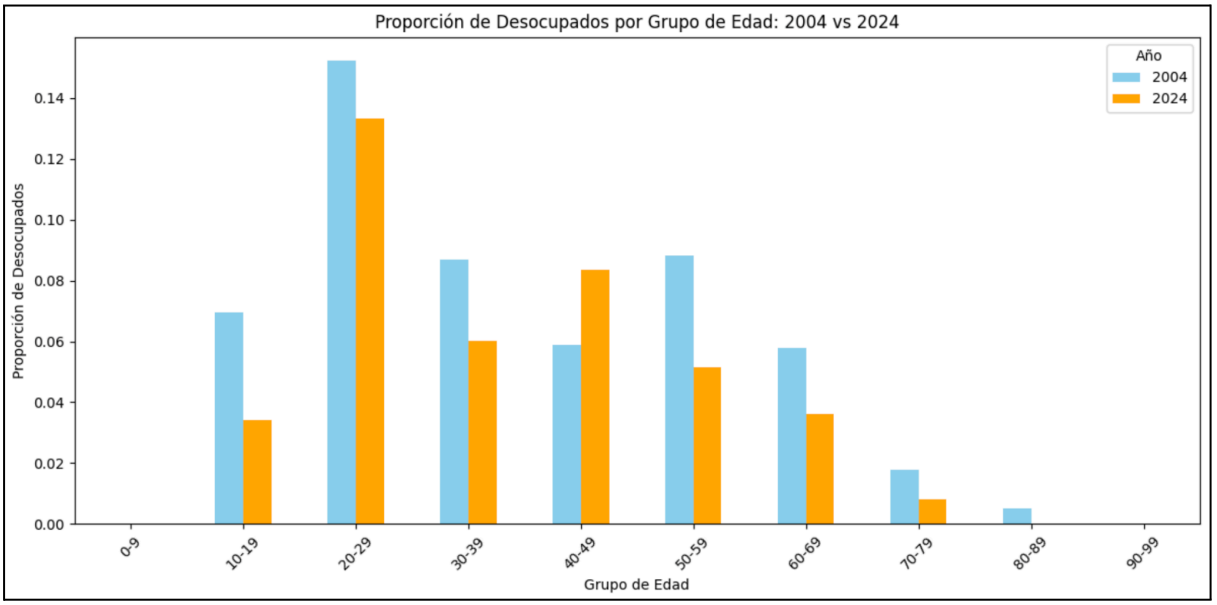
Se analizaron las proporciones de desocupados según el nivel educativo en 2004 y 2024. Los resultados mostraron una disminución general en la desocupación para la mayoría de los niveles educativos, aunque los niveles intermedios, como "Secundario incompleto", siguen concentrando la mayor proporción de desocupados.



Punto 6b (Desocupados por grupo de edad):

Se agruparon las edades en intervalos de 10 años para analizar la proporción de desocupados en cada grupo. La mayoría de los grupos de edad mostraron una disminución

en la desocupación entre 2004 y 2024, con la excepción del grupo de 40-49 años, donde la desocupación aumentó ligeramente.

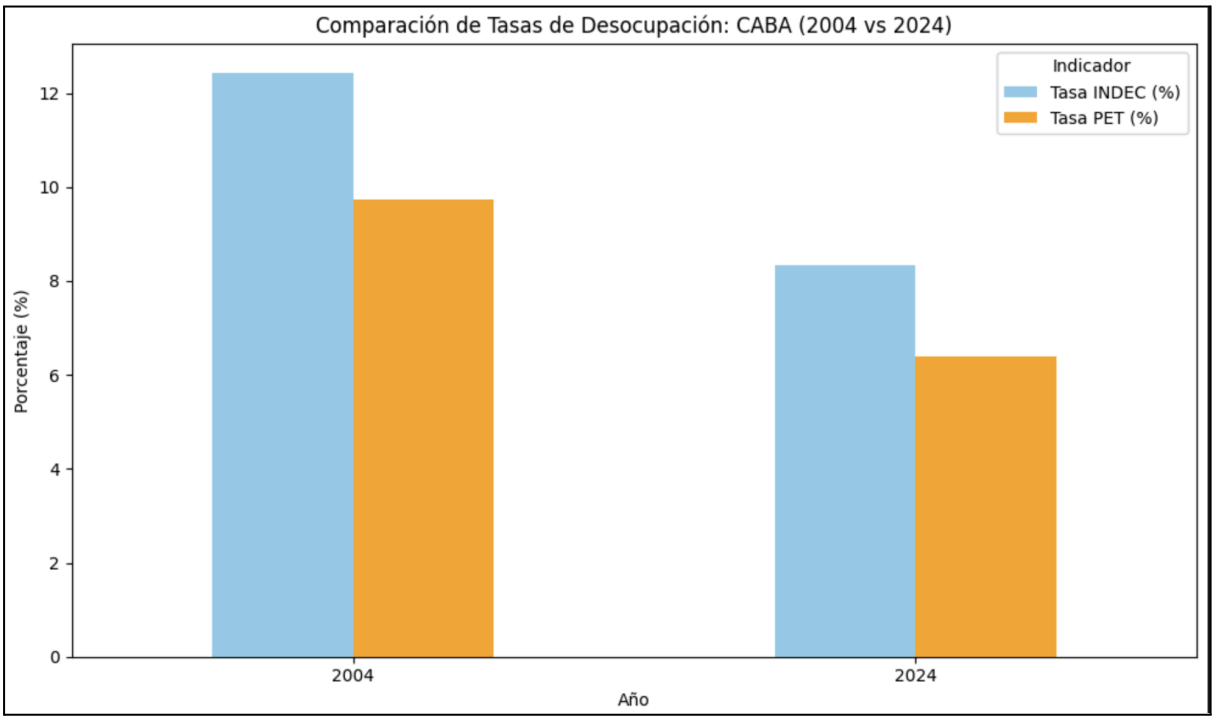


Punto 7

En el Punto 7, se calcularon dos tasas de desocupación para CABA en 2004 y 2024, utilizando dos enfoques distintos para ofrecer una visión comparativa de la dinámica laboral:

1. **Tasa de desocupación según el INDEC:**
   Esta tasa se calculó como el porcentaje de personas desocupadas sobre la **PEA** (Población Económicamente Activa), reflejando la proporción de desocupados entre quienes participan activamente en el mercado laboral. Este enfoque es el estándar utilizado por organismos oficiales para medir la desocupación.
2. **Tasa de desocupación alternativa respecto a la PET:**
   En esta tasa, se consideró el porcentaje de personas desocupadas sobre la **PET** (Población en Edad para Trabajar). Este enfoque amplía el análisis al incluir a toda la población en edad de trabajar, independientemente de su participación en el mercado laboral, captando dinámicas relacionadas con la inactividad.

Ambas tasas mostraron una disminución significativa en 2024 en comparación con 2004, indicando una mejora general en la situación laboral. Sin embargo, la tasa INDEC fue consistentemente más alta, ya que se calcula solo sobre la PEA, mientras que la tasa PET, al incluir a los inactivos, reduce la proporción relativa de desocupados.



Parte II: Clasificación

Para predecir si una persona está desocupada o no, utilizamos diferentes métodos de clasificación supervisada sobre las características individuales disponibles en la base.

División de la muestra

Se partió la base 'respondieron' para cada año, usando train_test_split con un 70% de los datos para entrenamiento y estableciendo una semilla (random state) de 101. Se removieron las variables PEA y ESTADO por estar directamente relacionadas con la condición de desocupación.

Modelos implementados

Se implementaron cuatro modelos de clasificación:

- Regresión logística
- Análisis discriminante lineal (LDA)
- K vecinos más cercanos (KNN) con k=3
- Naive Bayes

Para cada uno se calculó la matriz de confusión, la curva ROC, y los valores de AUC y accuracy. Sin embargo, dado el marcado desbalance en nuestros datos (mucho menor proporción de desocupados que no desocupados), fue fundamental considerar métricas adicionales como la sensibilidad (proporción de desocupados correctamente identificados) y la precisión (proporción de aciertos cuando el modelo predice que alguien está desocupado).

Comparación de resultados

En 2004, tanto la regresión logística como el Naive Bayes mostraron excelentes resultados con sensibilidades superiores al 95% y precisiones cercanas al 66%. El KNN logró un desempeño intermedio con una sensibilidad del 41% y una precisión del 62%, mientras que el LDA mostró dificultades significativas identificando apenas el 3% de los desocupados.

Para 2024, la regresión logística se destacó claramente como el mejor método, mejorando incluso su desempeño anterior con una sensibilidad del 96% y una precisión del 84%. El Naive Bayes mostró un comportamiento curioso: mantuvo una excelente sensibilidad del 99% pero su precisión cayó dramáticamente al 8%, indicando que generó muchos falsos positivos. El KNN y LDA también vieron deteriorado su desempeño en este período.

Cuando se analiza la base completa sin separar por años, la regresión logística mantiene un desempeño robusto con una sensibilidad del 87% y una precisión del 70%, sugiriendo que es el método más confiable y estable a través del tiempo.

Predicción

Utilizando la regresión logística, que mostró el mejor desempeño en ambos años, se predijo sobre la base ‘no_respondieron’. De los 51 casos analizados, el modelo clasificó a 36 personas como desocupadas (70%). Este resultado contrasta fuertemente con la proporción real de desocupados observada en la base ‘respondieron’, que es de apenas 6.2%.

Si bien excluimos PEA y ESTADO por su relación directa con la desocupación (incluirlas sería equivalente a revelar la respuesta al modelo), el caso de P21 (ingresos) nos presentó un dilema: al examinar los datos vemos que en la base 'no_respondieron', la variable P21 tiene valores de 0 para todas las observaciones. La decisión de mantener P21 en el modelo muestra sus limitaciones: el modelo predice una tasa de desocupación del 70% cuando incluimos esta variable, pero al eliminarla, el modelo se vuelve incapaz de identificar desocupados en la base de test, prediciendo 0% de desocupación. Esto sugiere que P21 es una variable crucial para la identificación de desocupados, sin la cual el modelo pierde toda su capacidad predictiva.

La importancia de P21 se evidencia en que, al tener valores de 0 en la base 'no_respondieron', el modelo clasifica a la mayoría como desocupados (70%), mientras que sin esta variable el modelo no logra identificar ningún caso de desocupación en la base de test, mostrando que los datos de ingresos son fundamentales para la capacidad predictiva del modelo.