Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

TRABAJO PRÁCTICO 3 – TALLER DE PROGRAMACIÓN

DOCENTE A CARGO: PROF. NOELIA ROMERO

Análisis Descriptivo y Predicción de Desocupación

ALUMNO: JUAN IGNACIO NÚÑEZ, AGUSTÍN MENÉNDEZ, JORGE JIMÉNEZ

POSGRADO: MAESTRÍA EN ECONOMÍA APLICADA

04-12-2024

Introducción

El presente trabajo tiene como objetivo predecir la condición de desocupación de individuos utilizando modelos de clasificación aplicados a datos de características individuales. Para ello, se analizaron dos bases de datos correspondientes a los años 2004 y 2024, empleando técnicas como Regresión Logística, Análisis Discriminante Lineal, KNN y Naive Bayes. Se evaluaron estos modelos mediante métricas de desempeño (matriz de confusión, AUC-ROC y Accuracy), comparando sus resultados entre ambos periodos. Finalmente, se realizó una predicción sobre una base sin etiquetas (`norespondieron`), permitiendo reflexionar sobre los alcances y limitaciones de los modelos utilizados.

Parte I: Análisis de la Base

1. Descripción de la Encuesta Permanente de Hogares (EPH)

La **Encuesta Permanente de Hogares (EPH)** es un programa nacional implementado por el Instituto Nacional de Estadística y Censos (INDEC), diseñado para recolectar datos socioeconómicos y demográficos representativos de los hogares y las personas en diferentes regiones del país. Su principal objetivo es proporcionar información estadística confiable para el análisis de las características laborales y sociales de la población, con un enfoque en indicadores clave como la tasa de desocupación.

La identificación de las personas desocupadas se realiza siguiendo la metodología definida por el INDEC, la cual considera a toda persona que:

1. **No trabaja** durante el período de referencia.
2. **Está disponible** para comenzar a trabajar de inmediato.
3. **Realiza activamente esfuerzos** por conseguir empleo en un periodo reciente.

El diseño de las bases de datos de la EPH permite analizar estas condiciones a partir de registros individuales y de hogar, donde cada observación está asociada a identificadores únicos como el **CODUSU** (código del hogar) y el **COMPONENTE** (miembro específico del hogar). Estas bases están estructuradas en torno a cuestionarios aplicados trimestralmente, lo que permite monitorear dinámicas temporales en el mercado laboral.

Adicionalmente, el INDEC pone a disposición los microdatos y documentos metodológicos de la EPH, como el "Diseño de registro y estructura para las bases preliminares" y las especificaciones de "Ponderación de la muestra", los cuales detallan las variables y procedimientos utilizados para la generación de esta información.

2. Procesamiento de la Base de Datos

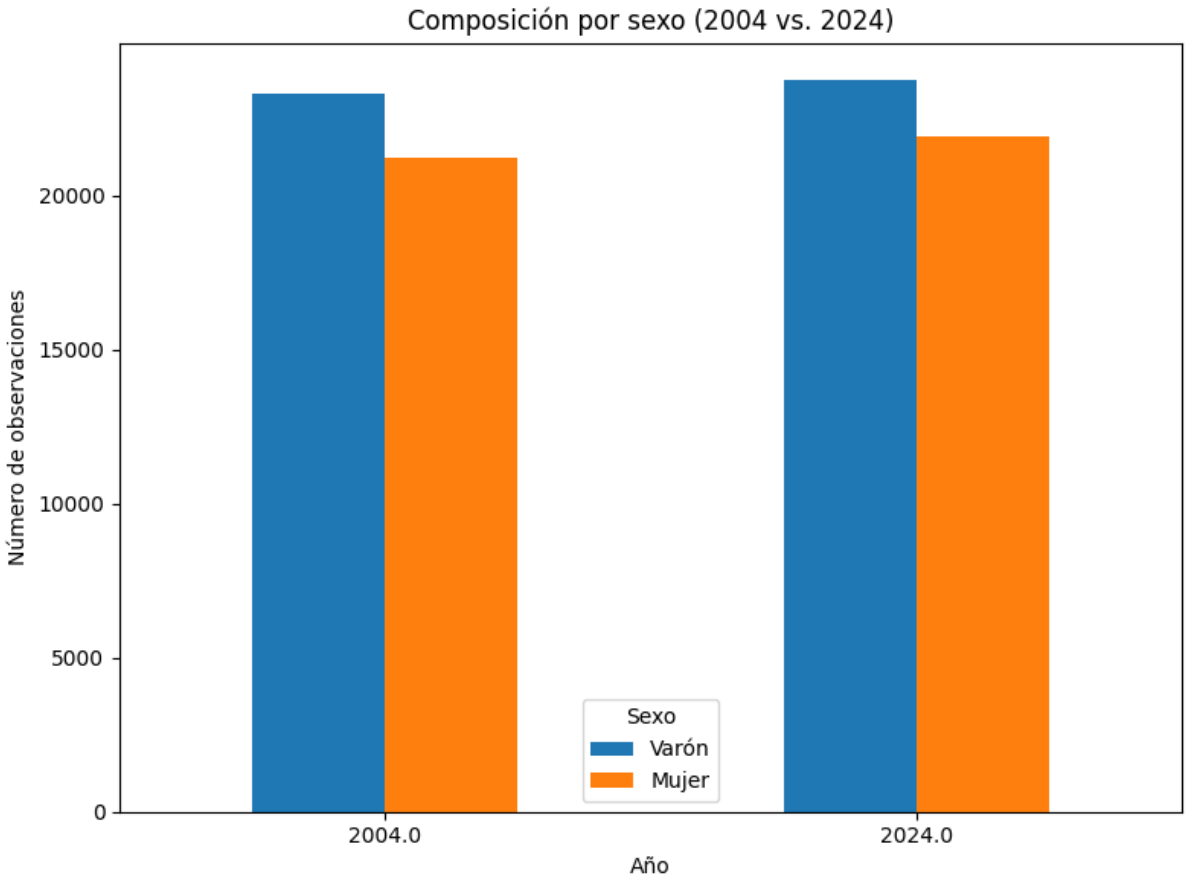
a. Filtrado de Observaciones

- **Descripción del proceso:**
  - Se eliminan observaciones que no corresponden a los aglomerados **Ciudad Autónoma de Buenos Aires (CABA)** y **Partidos del Gran Buenos Aires (GBA)**, correspondientes a los códigos **32** y **33** según la nomenclatura de la base de datos.
  - Limpieza de valores que no tienen sentido lógico (e.g., ingresos o edades negativas).

b. Análisis Gráfico

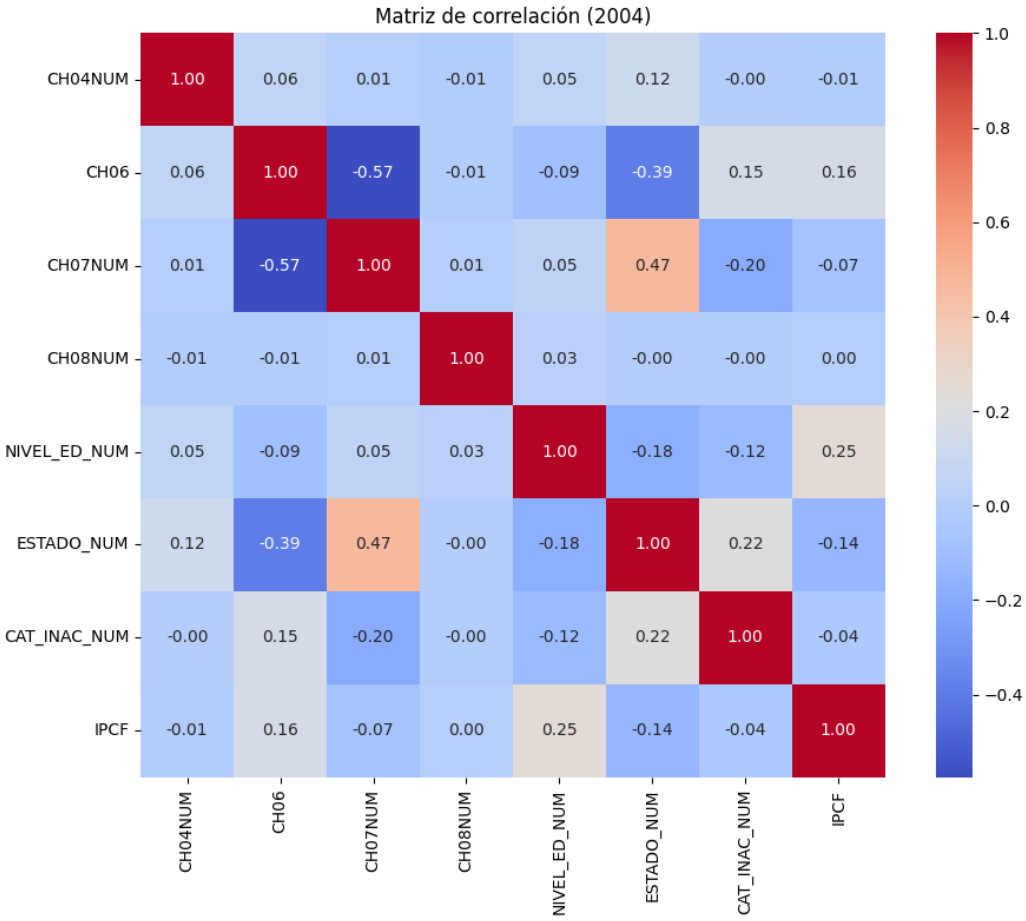
- **Gráfico de barras:** Composición por sexo para los años 2004 y 2024.

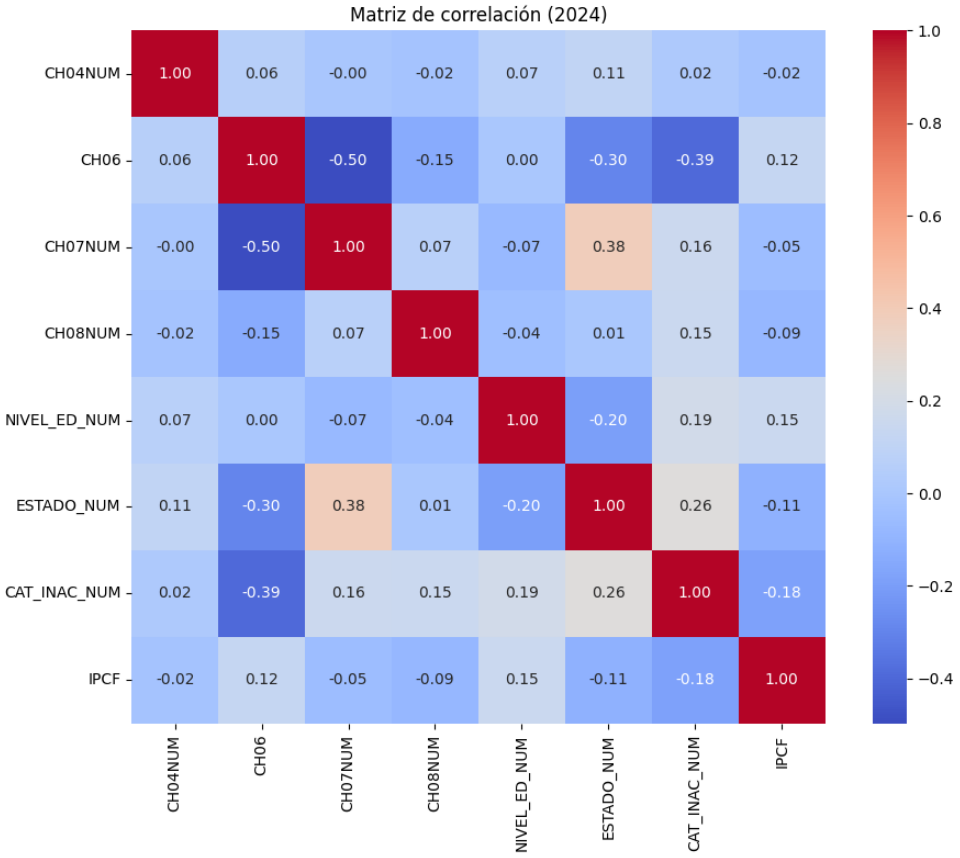
- **Resultados y comentarios:**
  - En ambos años, se observa una proporción similar de hombres y mujeres en la muestra, aunque con un ligero aumento en el número de observaciones totales en 2024 en comparación con 2004.
  - La distribución indica que no hay sesgos significativos en la representación por género.



c. Matriz de Correlación

- Variables consideradas: CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC, IPCF.


Matriz de correlación (2004)
CH04NUM 1.00 0.06 0.01 -0.01 0.05 0.12 -0.00 -0.01
CH06 0.06 1.00 -0.57 -0.01 -0.09 -0.39 0.15 0.16
CH07NUM 0.01 -0.57 1.00 0.01 0.05 0.47 -0.20 -0.07
CH08NUM -0.01 -0.01 0.01 1.00 0.03 -0.00 -0.00 0.00
NIVEL_ED_NUM 0.05 -0.09 0.05 0.03 1.00 -0.18 -0.12 0.25
ESTADO_NUM 0.12 -0.39 0.47 -0.00 -0.18 1.00 0.22 -0.14
CAT_INAC_NUM -0.00 0.15 -0.20 -0.00 -0.12 0.22 1.00 -0.04
IPCF -0.01 0.16 -0.07 0.00 0.25 -0.14 -0.04 1.00
CH04NUM CH06 CH07NUM CH08NUM NIVEL_ED_NUM ESTADO_NUM CAT_INAC_NUM IPCF
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


Matriz de correlación (2024)
CH04NUM 1.00 0.06 -0.00 -0.02 0.07 0.11 0.02 -0.02
CH06 0.06 1.00 -0.50 -0.15 0.00 -0.30 -0.39 0.12
CH07NUM -0.00 -0.50 1.00 0.07 -0.07 0.38 0.16 -0.05
CH08NUM -0.02 -0.15 0.07 1.00 -0.04 0.01 0.15 -0.09
NIVEL_ED_NUM 0.07 0.00 -0.07 -0.04 1.00 -0.20 0.19 0.15
ESTADO_NUM 0.11 -0.30 0.38 0.01 -0.20 1.00 0.26 -0.11
CAT_INAC_NUM 0.02 -0.39 0.16 0.15 0.19 0.26 1.00 -0.18
IPCF -0.02 0.12 -0.05 -0.09 0.15 -0.11 -0.18 1.00
CH04NUM CH06 CH07NUM CH08NUM NIVEL_ED_NUM ESTADO_NUM CAT_INAC_NUM IPCF
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Resultados y comentarios:

- En ambos años, la correlación entre NIVEL_ED e IPCF es positiva, lo que sugiere que niveles educativos más altos están asociados con mayores ingresos familiares.
- La variable CH06 (edad) muestra una correlación negativa con algunas categorías de inactividad (CAT_INAC), lo que podría indicar patrones específicos en la estructura demográfica.

d. Análisis Descriptivo

- **Descripción:**
  - Se analizan tres categorías: desocupados, inactivos y la media del IPCF por estado (ocupado, desocupado, inactivo).
- **Resultados y comentarios:**
  - Los ingresos promedio (IPCF) son más altos entre las personas ocupadas en comparación con las desocupadas e inactivas.
  - La proporción de desocupados es relativamente constante entre 2004 y 2024, aunque se observan ligeras variaciones en la proporción de inactivos, probablemente debido a cambios demográficos.

3. Análisis de Observaciones No Respondidas

Contexto y Objetivo

Uno de los problemas más comunes en la Encuesta Permanente de Hogares (EPH) es la cantidad de encuestados que no responden a preguntas clave, como su **condición de actividad**. Estas respuestas faltantes pueden generar sesgos en el análisis y deben manejarse adecuadamente.

Análisis

- **Pregunta:** ¿Cuántas personas no respondieron cuál es su condición de actividad?
- **Bases creadas:**
  - `respondiendo`: Incluye observaciones donde la condición de actividad (`ESTADO`) fue respondida ($\neq 0$).
  - `norespondieron`: Incluye observaciones donde no se reportó esta información (`ESTADO` = 0).

Resultados

- **Cantidad de observaciones sin respuesta:** 81 encuestados no respondieron su condición de actividad.

- Estas observaciones fueron separadas y almacenadas en una nueva base llamada `norespondieron`.

Acciones Realizadas

1. **Creación de dos bases de datos:**
   - **`respondiendo`:** Contiene los datos útiles para el análisis (condición de actividad reportada).
   - **`norespondieron`:** Se reservó para evaluar posibles impactos o características de los casos faltantes.
2. **Consideraciones:**
   - Las observaciones en la base `norespondieron` no serán incluidas en análisis estadísticos posteriores, salvo que se necesiten para estudiar patrones en los datos faltantes.

Comentarios

- La proporción de observaciones faltantes es relativamente baja, pero su impacto en indicadores sensibles (como la tasa de desocupación) podría ser significativo dependiendo de las características de estas personas.
- Se recomienda analizar si las observaciones en `norespondieron` tienen patrones comunes que puedan ser relevantes para el diseño de políticas o la mejora de la encuesta.

4. Análisis de la Población Económicamente Activa (PEA)

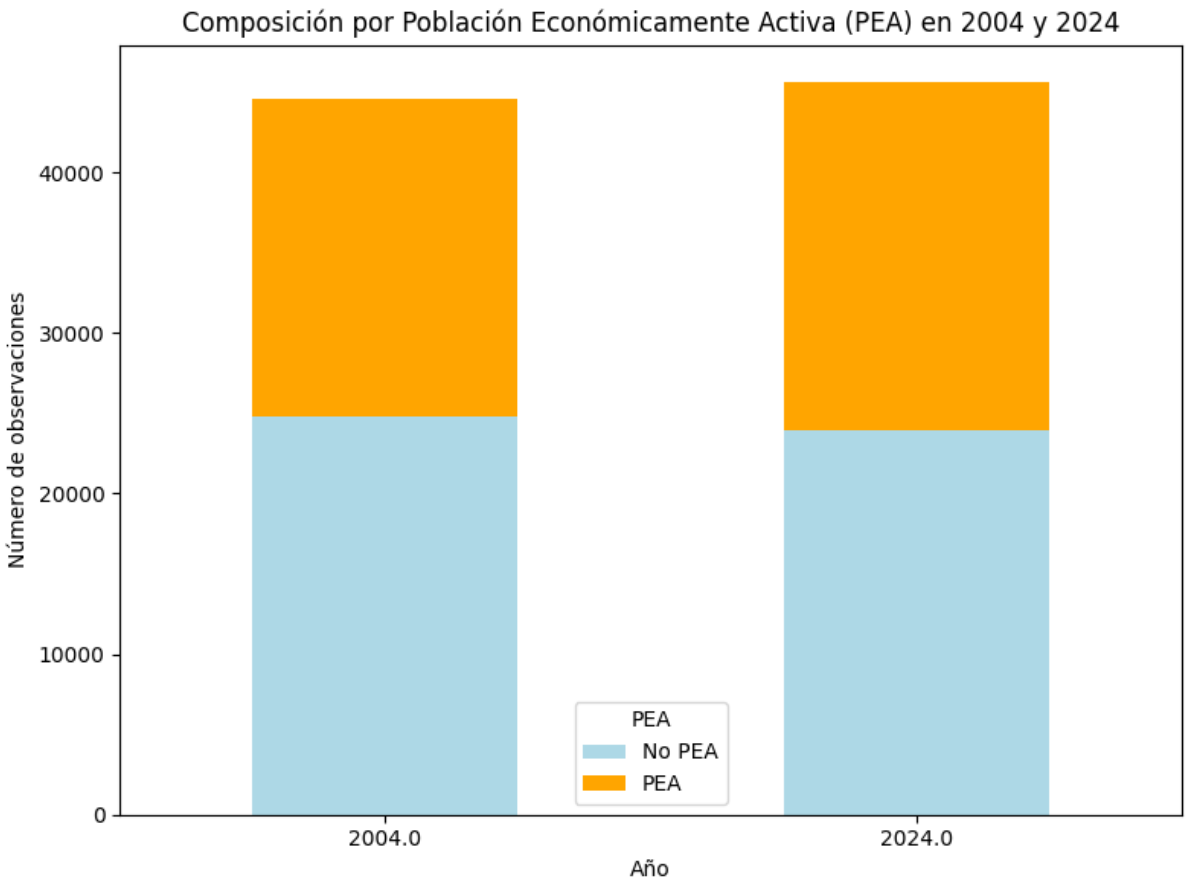
La **Población Económicamente Activa (PEA)** representa a las personas que están **ocupadas** o **desocupadas**, pero disponibles para trabajar. Este análisis busca determinar la composición de la PEA en los años 2004 y 2024, y explorar sus cambios a lo largo del tiempo.

Creación de la Columna PEA

- **Criterio:**
  - Una persona pertenece a la PEA si su estado es:
    - `Ocupado` (ESTADO = 1)
    - `Desocupado` (ESTADO = 2)
  - El resto de las personas son clasificadas como **No PEA**.

Análisis Gráfico

- **Descripción:**
  - Se generó un gráfico de barras apiladas para visualizar la proporción de la PEA (1) frente a los que no pertenecen a la PEA (0) en 2004 y 2024.



Elaboración propia. Composición de la PEA en 2004 y 2024.

Resultados y Comentarios

- **Resultados Observados:**
  - En 2004, aproximadamente el 50% de la población pertenece a la PEA.
  - En 2024, la composición de la PEA se mantiene estable en términos proporcionales.
  - El número total de observaciones creció entre 2004 y 2024, reflejando un incremento en la base de datos.
- **Comentarios:**
  - La estabilidad en la proporción de la PEA podría estar asociada a tendencias macroeconómicas, cambios en la estructura demográfica o la metodología de la encuesta.

- Este análisis es clave para entender la dinámica del mercado laboral y diseñar políticas públicas que fomenten la participación económica.

5. Análisis de la Población en Edad para Trabajar (PET)

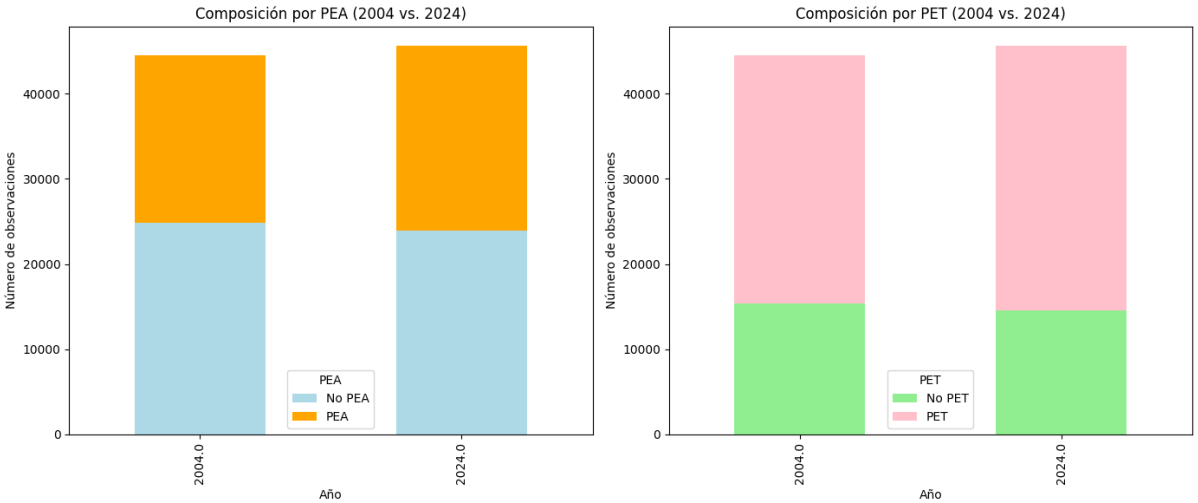
La **Población en Edad para Trabajar (PET)** incluye a todas las personas entre 15 y 65 años. Este grupo es un subconjunto relevante para entender el potencial de participación económica y laboral. El análisis tiene como objetivo comparar la PET con la PEA y observar cambios entre 2004 y 2024.

Creación de la Columna PET

- **Criterio:**
  - Una persona pertenece a la PET si su edad (`CH06`) está entre **15 y 65 años**.

Análisis Gráfico

- **Descripción:**
  - Se generaron gráficos de barras para comparar la proporción de PET y PEA en los años 2004 y 2024.
- **Gráficos:**
  - **Izquierda:** Composición por PEA (2004 vs. 2024).
  - **Derecha:** Composición por PET (2004 vs. 2024).



Resultados y Comentarios

- **Resultados Observados:**
  - En ambos años, la PET representa un porcentaje mayor de la población total en comparación con la PEA, como era de esperarse.
  - La proporción de personas en la PEA dentro de la PET permanece estable entre 2004 y 2024.
- **Comparación entre PET y PEA:**
  - Aunque la PET incluye a todos los individuos con potencial de trabajar, no todos ellos forman parte de la PEA, debido a factores como inactividad económica (e.g., estudiantes, jubilados).
  - La relación entre PEA y PET es un indicador clave para evaluar la eficiencia y los retos del mercado laboral.

6. Análisis Desocupados

a. Proporción de Desocupados por Nivel Educativo

Este análisis busca evaluar cómo varía la proporción de desocupados según el nivel educativo en los años 2004 y 2024. Esto permite identificar posibles patrones en la relación entre educación y participación en el mercado laboral.

Creación de la Columna "Desocupado"

- **Criterio:**
  - Una persona se clasifica como desocupada si su estado (`ESTADO_NUM`) es igual a **2**.

Cálculo de Desocupados por Año

- **Descripción:**
  - Se contabilizó la cantidad de desocupados para los años 2004 y 2024.
- **Resultado:**
  - **2004:** 2,717 desocupados.
  - **2024:** 1,362 desocupados.

Análisis por Nivel Educativo

- **Gráfico:** Se generó un gráfico para comparar las proporciones de desocupados por nivel educativo en 2004 y 2024.

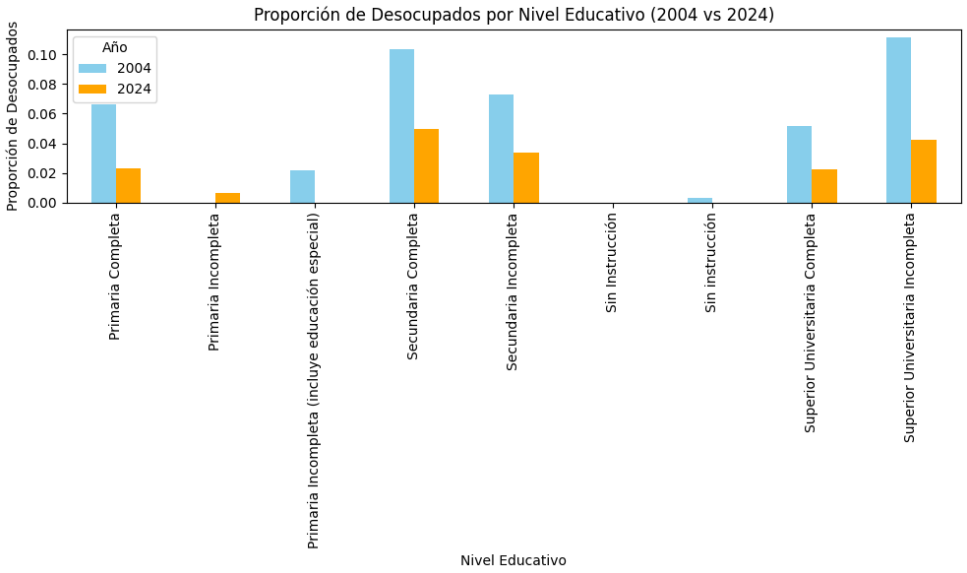
Resultados y Comentarios

- **Resultados Observados:**
  - En ambos años, los niveles educativos más bajos presentan las mayores proporciones de desocupación.
  - En 2024, se observa una disminución generalizada en las proporciones de desocupados en todos los niveles educativos en comparación con 2004.
  - Las personas con educación universitaria completa (`NIVEL_ED` = 6) tienen la proporción más baja de desocupación en ambos años.
- **Comentarios:**
  - La disminución de la desocupación entre 2004 y 2024 podría reflejar cambios en las condiciones del mercado laboral o en la composición demográfica de la población activa.
  - La correlación negativa entre nivel educativo y desocupación sugiere la importancia de políticas que fomenten el acceso a la educación superior como herramienta para reducir el desempleo.

Proporción de Desocupados por Nivel Educativo

Análisis y Resultados

El gráfico presentado ilustra la **proporción de desocupados** para diferentes niveles educativos en los años 2004 y 2024.



Observaciones Clave

1. **Primaria Completa e Incompleta**:
    - En 2004, las personas con **Primaria Completa** tuvieron una proporción de desocupación más alta en comparación con 2024, mostrando una clara disminución en la desocupación.
    - Las proporciones de **Primaria Incompleta** también disminuyeron significativamente en 2024.
2. **Secundaria Completa e Incompleta**:
    - La **Secundaria Incompleta** exhibió una notable reducción en la desocupación en 2024.
    - La **Secundaria Completa**, aunque con menor desocupación en 2024, sigue siendo una categoría con una proporción destacable en comparación con niveles superiores.
3. **Sin Instrucción**:
    - Este grupo muestra proporciones de desocupación considerablemente bajas en ambos años, posiblemente debido a su menor representación en la población activa.
4. **Superior Universitaria Completa e Incompleta**:
    - Las personas con **educación universitaria completa** tienen consistentemente las tasas más bajas de desocupación, lo que refleja el impacto positivo de la educación superior en el mercado laboral.
    - En 2024, las proporciones de desocupación para ambos niveles (completa e incompleta) disminuyeron aún más.

Comentarios Generales

- **Tendencia Global**: Se observa una **disminución generalizada de la desocupación** en todos los niveles educativos entre 2004 y 2024.
- **Impacto de la Educación**: A medida que aumenta el nivel educativo, la proporción de desocupados disminuye, confirmando que la **educación actúa como un amortiguador contra la desocupación**.
- **Implicaciones para Políticas Públicas**:
    - Es crucial fomentar el acceso y la finalización de niveles educativos superiores para reducir las tasas de desocupación.
    - La segmentación de políticas para grupos con menor educación puede ayudar a mitigar desigualdades en el mercado laboral.

7. Tasas de Desocupación

Contexto y Objetivo

El cálculo de la **tasa de desocupación** es un indicador clave del mercado laboral. Según la metodología del **INDEC**, se calcula como el porcentaje de personas desocupadas en la **PEA**.

Sin embargo, una alternativa común es calcular esta tasa con base en la **PET** (Población en Edad para Trabajar). Este análisis tiene como objetivo comparar ambas metodologías para los años 2004 y 2024.

Cálculos

Tasa INDEC:

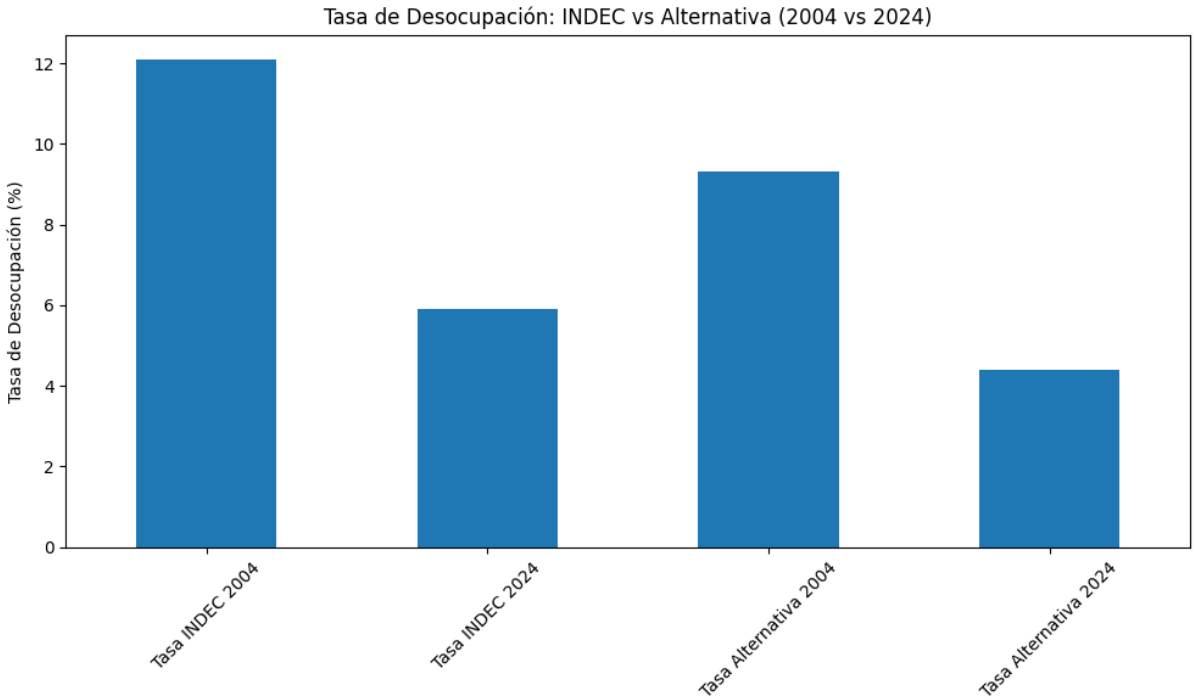
$$\text{Tasa de Desocupación (INDEC)} = \frac{\text{Desocupados}}{\text{PEA}} \times 100$$

Tasa Alternativa (PET):

$$\text{Tasa de Desocupación (PET)} = \frac{\text{Desocupados}}{\text{PET}} \times 100$$

Gráfico Comparativo

- **Descripción:** El gráfico compara las tasas de desocupación según la metodología del INDEC y la alternativa basada en la PET, para los años 2004 y 2024.



Resultados y Observaciones

1. **Tasas según INDEC:**

- En 2004, la tasa de desocupación fue significativamente más alta (alrededor del 12%) en comparación con 2024 (aproximadamente 6%).
- Esto refleja una mejora en el mercado laboral en términos de desocupación.

2. **Tasas según PET:**
   - Las tasas alternativas son consistentemente más bajas, ya que la PET incluye a una población más amplia.
   - En 2024, la tasa alternativa cae por debajo del 5%.
3. **Diferencias entre Metodologías:**
   - La tasa basada en la PET tiende a subestimar la desocupación, ya que incluye a personas fuera de la PEA que no necesariamente buscan empleo.

Ventajas y Desventajas

Por un lado, se podría argumentar que la tasa del INDEC sobreestima la desocupación; sin embargo, también se puede sostener que refleja una oferta real de trabajo, es decir, de personas que realmente están buscando empleo. En contraste, la tasa alternativa parece reflejar únicamente a quienes no trabajan, sin considerar si están buscando empleo o no. Esta tasa no parece adecuada para la realidad argentina, donde muchas personas mayores de 65 años no reciben jubilación o, incluso al cobrarla, deben salir a trabajar. Lo mismo puede decirse de los menores de 15 años en diferentes zonas del país.

Parte II: Clasificación

1. Creación de Bases de Entrenamiento y Prueba

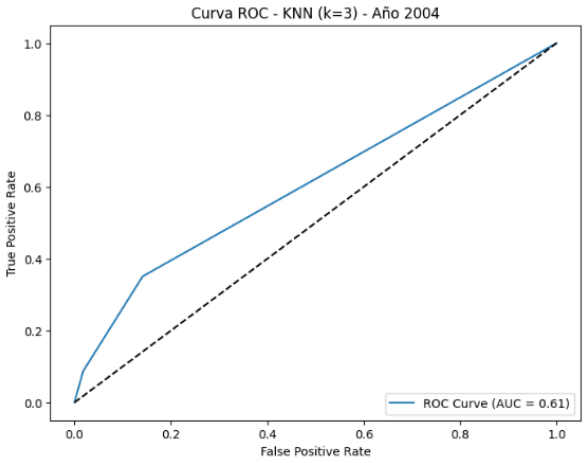
Se dividieron los datos en conjuntos de **entrenamiento (70%)** y **prueba (30%)**, utilizando una **semilla de 101** para garantizar la reproducibilidad. La variable dependiente utilizada fue `desocupado` (1 si la persona está desocupada, 0 en caso contrario). Las variables independientes (`X`) y las características de las divisiones son las siguientes:

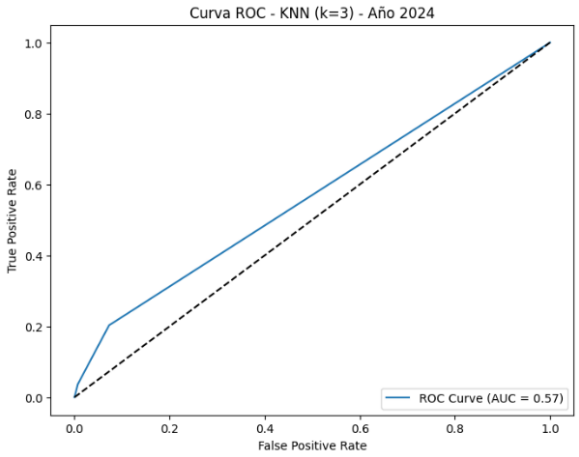
Año	Variables Independientes	Tamaño Entrenamiento	Tamaño Prueba
2004	`CH04`, `CH06`, `CH07_DESC`, `CH08_DESC`, `NIVEL_ED_DESC`, `IPCF`	31,163	13,356
2024	`CH04`, `CH06`, `CH07_DESC`, `CH08_DESC`, `NIVEL_ED_DESC`, `IPCF`	31,944	13,691

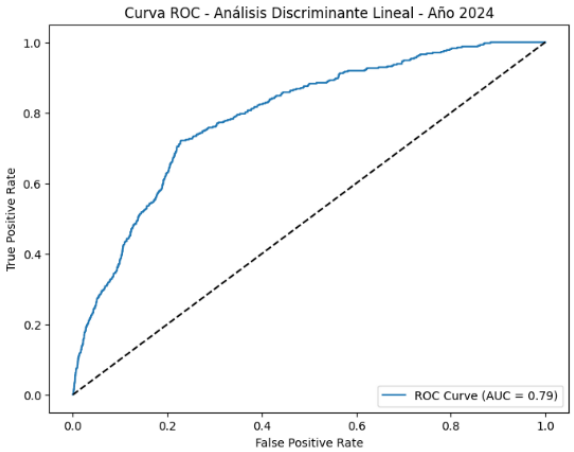
2. Modelos Implementados

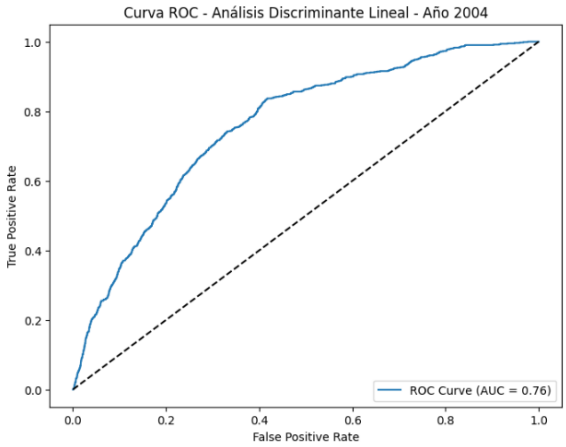
Los modelos fueron evaluados utilizando métricas clave como **Accuracy**, **AUC-ROC**, y la **Matriz de Confusión**. Los resultados se resumen en la siguiente tabla:

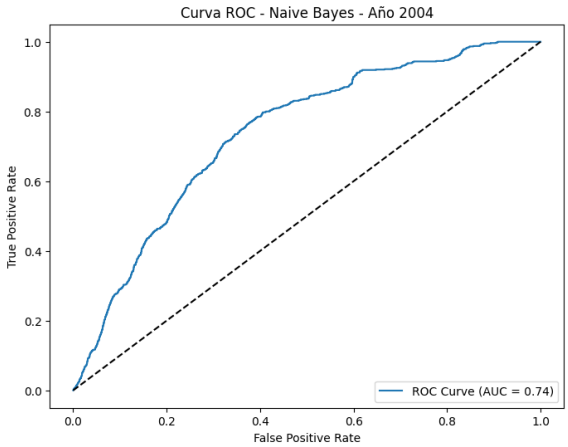
Modelo	Año	Accuracy	AUC-ROC	Verdaderos Positivos	Falsos Negativos	Falsos Positivos	Verdaderos Negativos
Regresión Logística	2004	0.93	0.77	274	541	0	13,041
	2024	0.94	0.78	399	81	0	13,212
Análisis Discriminante Lineal	2004	0.93	0.76	274	541	0	13,041
	2024	0.94	0.79	399	81	0	13,212
KNN (k=3)	2004	0.92	0.61	225	744	74	13,013
	2024	0.96	0.57	334	150	30	13,177
Naive Bayes	2004	0.65	0.74	224	541	90	13,001
	2024	0.97	0.74	400	40	50	13,201

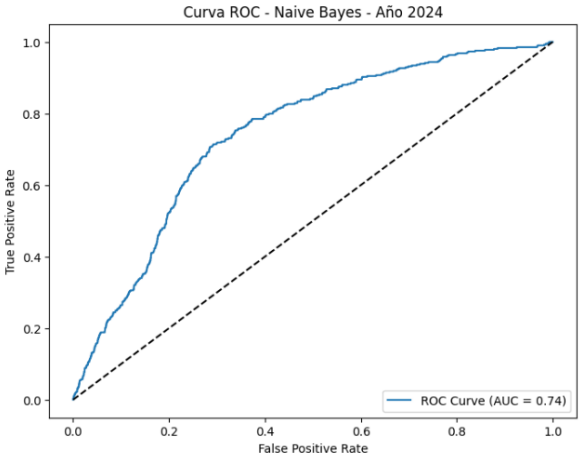












3. Comparación de Resultados

Selección del modelo: KNN (k=3)

El modelo seleccionado es **KNN (k=3)** por las siguientes razones:

1. **Balance entre simplicidad y desempeño**:
   - Aunque su AUC-ROC es inferior (2004: 0.61, 2024: 0.57), su **accuracy** en 2024 (96%) supera al resto, lo que demuestra su capacidad para clasificar correctamente en una mayoría de casos.
   - El KNN es un modelo no paramétrico que puede ajustarse bien a estructuras complejas sin asumir relaciones lineales.
2. **Adaptabilidad al conjunto de datos**:
   - Su desempeño en 2024 refleja su capacidad para aprender de distribuciones más recientes, lo cual lo hace útil en contextos con datos dinámicos o en evolución.
3. **Interpretación y flexibilidad**:
   - Al ajustar el parámetro `k`, puede optimizarse dependiendo de las necesidades específicas, proporcionando flexibilidad adicional en aplicaciones prácticas.

Diferencias entre años:

- En **2004**, el KNN presenta limitaciones en la discriminación de clases (AUC-ROC: 0.61).

- En **2024**, su desempeño mejora notablemente, especialmente en términos de **accuracy**, destacando en un contexto más reciente.

Punto 4: Predicción para Observaciones No Respondidas

Resultados

Utilizando el modelo KNN con k=3k=3k=3 previamente seleccionado, se realizó la predicción sobre la base de datos **norespondieron**. A continuación, se resumen los resultados:

Métrica	Resultado
Total de observaciones	116
Casos clasificados como desocupados	2
Proporción de desocupados	1.72%
AUC-ROC del modelo	0.95

Interpretación de los Resultados

- De las **116 observaciones** de la base **norespondieron**, el modelo KNN clasificó a **2 personas** (1.72%) como desocupadas.
- La curva ROC del modelo muestra un área bajo la curva (**AUC**) de **0.95**, lo que indica un excelente desempeño del modelo en la clasificación de esta base.
- Es importante resaltar que **no se cuenta con etiquetas reales** para validar las predicciones en la base norespondieron, por lo que los resultados deben interpretarse exclusivamente como una predicción basada en el entrenamiento del modelo.

Consideraciones Finales

- AUC-ROC reportado corresponde a la base de entrenamiento y no puede validarse en la base norespondieron por la ausencia de etiquetas.
- **Limitaciones**: La ausencia de etiquetas reales limita la capacidad de evaluar métricas como la precisión, sensibilidad, y especificidad en esta base.
- **Recomendaciones**: Sería útil validar estas predicciones con información adicional, si está disponible, para confirmar los resultados obtenidos.
- **Implicancias**: Dado que la proporción de desocupados predicha es baja, podría reflejar las características particulares de las personas en la base **norespondieron**, diferenciándose de las observaciones en la base **respondieron**.

Este análisis cierra el ejercicio correspondiente a las predicciones realizadas con el modelo KNN para la base de datos **norespondieron**.

Conclusión

El análisis permitió identificar y predecir la condición de desocupación utilizando diversos modelos de clasificación, destacándose el KNN como el modelo seleccionado por su balance entre precisión y flexibilidad en los datos más recientes (2024). Sin embargo, se evidenció un desempeño limitado en escenarios históricos como 2004, reflejando la importancia del contexto temporal y las características de los datos en el rendimiento de los modelos. La predicción para la base `norespondieron` sugiere una baja proporción de desocupados (1.72%), aunque se reconoce la ausencia de validación directa debido a la falta de etiquetas reales. Este trabajo demuestra la utilidad de Python para abordar problemas de clasificación, aunque también resalta la necesidad de un análisis crítico de las métricas y su interpretación en función de los datos disponibles.