

Muy buen trabajo!  
El código está muy bien.

NOTA: 9,5

# Trabajo Práctico 3


## Machine Learning para Economistas

Marotta, Salischiker, Zapiola

15 de Noviembre, 2024


### Parte I: Analizando la base

#### Inciso 1

El INDEC<sup>1</sup> define a los individuos desocupados como aquellas personas sin empleo que buscan trabajo activamente y están disponibles para comenzar a trabajar. Para identificarlos, se sigue un proceso específico a partir del universo de la Encuesta Permanente de Hogares (EPH), que representa a 29.7 millones de personas. 

Primero, se distingue la población económicamente activa (PEA), que incluye tanto a los individuos que tienen una ocupación como a aquellos que, sin estar ocupados, están buscando activamente empleo y disponibles para trabajar. Dentro de la PEA, se consideran desocupados a los individuos que no tienen ocupación pero cumplen con los criterios de búsqueda activa y disponibilidad laboral. Es decir, individuos sin ocupación pero que no están buscando activamente trabajo, no están incluidos en la población de desocupados.

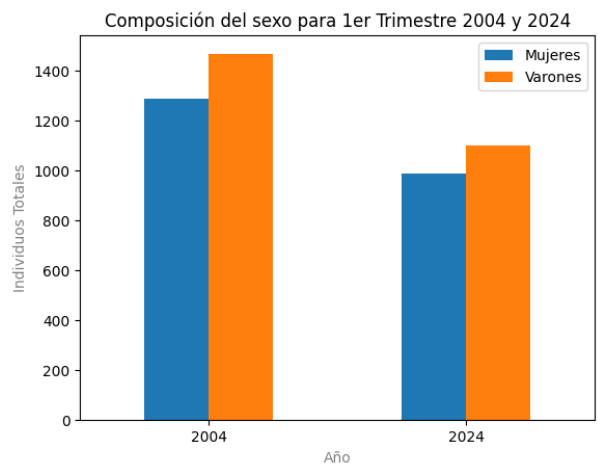
#### Inciso 2

- a) Ver *Python Script*.
- b) Las observaciones descartadas fueron aquellas con ingresos y edades negativos o con valores faltantes (*missings*). La razón detrás de esta decisión es que, para realizar un análisis confiable sobre el estado laboral de las personas, es fundamental trabajar con datos que sean coherentes y representativos de la realidad. Las observaciones con ingresos negativos o edades fuera del rango lógico pueden distorsionar los resultados, ya que no reflejan situaciones posibles dentro de la población estudiada. Asimismo, los datos faltantes dificultan la interpretación y puede llevar a conclusiones erróneas o poco precisas. 
- c) La Figura 1 muestra la composición por sexo para el primer trimestre de los años 2004 y 2024. En ambos años, la cantidad de hombres en el aglomerado del Gran Córdoba, según la EPH, es mayor que la de mujeres. Además, el número total de personas, tanto hombres como mujeres, es superior en 2004 en comparación con 2024.

---

<sup>1</sup>[https://www.indec.gob.ar/uploads/informesdeprensa/mercado\\_trabajo\\_eph\\_2trim2404BDC5E521.pdf](https://www.indec.gob.ar/uploads/informesdeprensa/mercado_trabajo_eph_2trim2404BDC5E521.pdf)

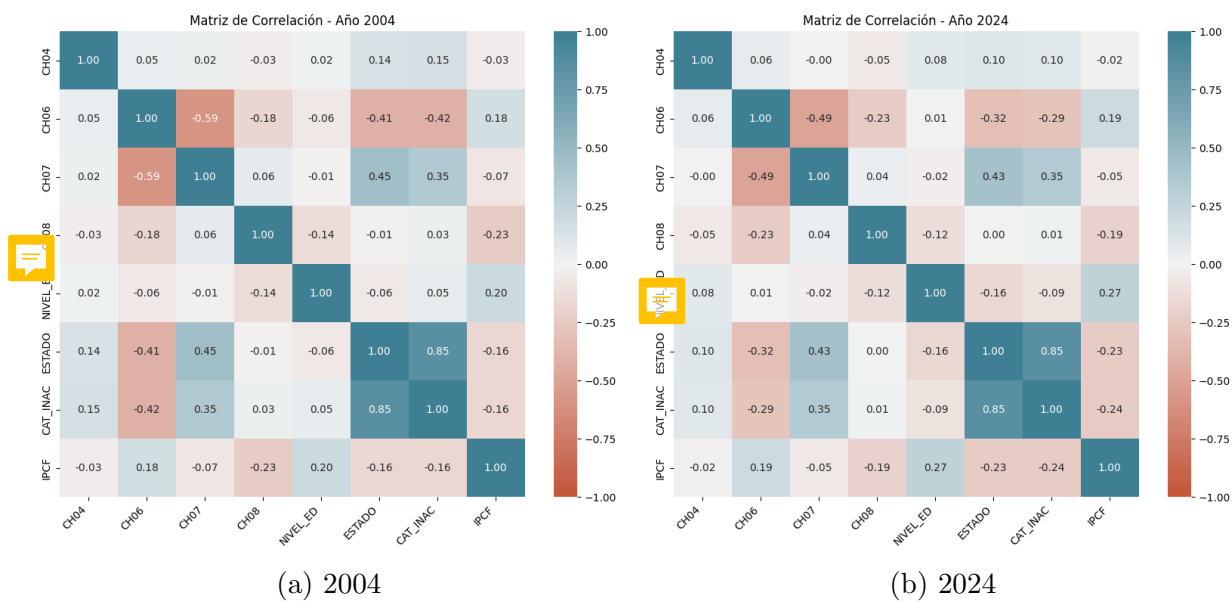
Figura 1: Composición del sexo para el primer trimestre del 2004 y 2024.



d) La Figura 2 presenta las matrices de correlación para el primer trimestre de los años 2004 y 2024. Las variables correlacionadas son las siguientes: sexo (CH04), edad (CH06), estado civil (CH07), cobertura médica (CH08), nivel educativo (NIVEL\_ED), condición de actividad (ESTADO), categoría de inactividad (CAT\_INAC), y monto del ingreso per cápita familiar (IPCF).

Al analizar visualmente las correlaciones entre variables categóricas ordinales (con un orden natural) y variables numéricas, se observa una correlación positiva entre la edad y el ingreso per cápita familiar, así como entre el nivel educativo y el ingreso per cápita familiar, tanto para los años 2004 como 2024. Sin embargo, en 2004 se observa una leve correlación negativa entre la edad y el nivel educativo, mientras que en 2024 esta correlación es positiva, aunque de magnitud muy baja.

Figura 2: Matrices de correlación para el 2004 y 2024.



e) En el primer trimestre de 2004, en la muestra hay 1056 ocupados con un ingreso per cápita familiar (IPCF) de \$ 364.46, 171 desocupados con un IPCF de \$181.21,

e 1065 inactivo con un IPCF de \$260.30.

En el primer trimestre de 2024, hay 959 ocupados con un IPCF de \$266394.39, 86 desocupados con un IPCF de \$134357.59, e 775 inactivos con un IPCF de \$187277.54.

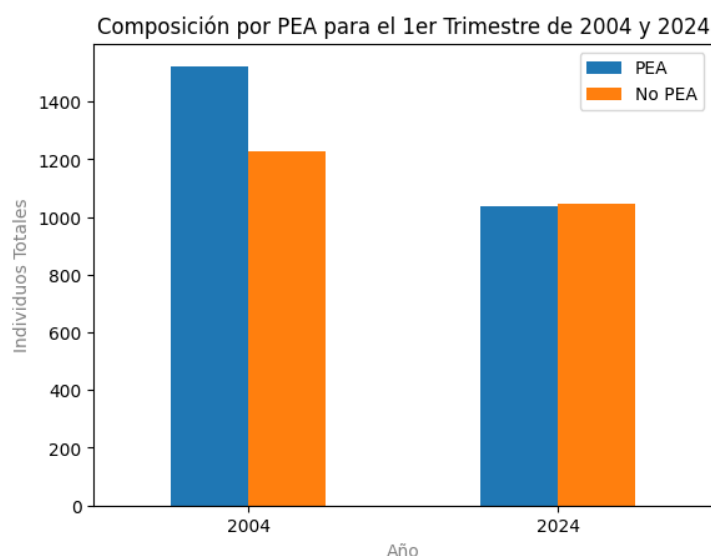
### Inciso 3

La cantidad de personas que no respondieron la pregunta sobre su condición de actividad (ESTADO) fue solamente una en el año 2004.

### Inciso 4

La Figura 3 muestra la composición por PEA (Población Económicamente Activa) para el primer trimestre del 2004 y 2024. La PEA incluye tanto a los individuos que tienen una ocupación como a aquellos que, sin estar ocupados, están buscando activamente empleo y disponibles para trabajar. Para el 2004, la cantidad de individuos que estaban económicamente activos era mayor a la que no, por una gran diferencia. Mientras que en el 2024, la cantidad de individuos que estaban económicamente activos era menor a los que si lo estaban.

Figura 3: Composición por PEA para el primer trimestre del 2004 y 2024.

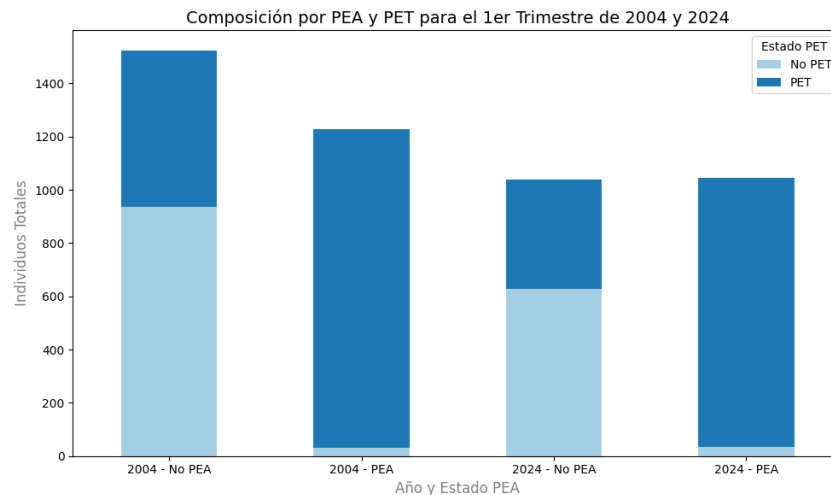


### Inciso 5

La Figura 4 muestra la composición de la población en edad de trabajar (PET) según su pertenencia a la población económicamente activa (PEA) para el primer trimestre de 2004 y 2024. En ambos años, la proporción de personas en edad de trabajar que están económicamente activas (PEA y PET) se mantiene casi sin cambios. Por otro lado, la cantidad de personas que no están en edad de trabajar y tampoco son económicamente activas (no PEA y no PET) es mayor que aquellas en edad de trabajar pero no activas económicamente (no PEA y PET) en ambos períodos. Sin embargo, la proporción de

personas en edad de trabajar que no participan en la actividad económica (no PEA y PET) es más alta en 2004 en comparación con 2024.

Figura 4: Composición del PET dado si pertenece al PEA para el primer trimestre del 2004 y 2024.



## Inciso 6

En el aglomerado de Gran Córdoba, en el primer trimestre de 2004, un 6.63 % de los individuos esta desocupado (171 individuos). Mientras que en el primer trimestre de 2024, un 4.30 % de los individuos esta desocupado (86 individuos).

- La Tabla 1 presenta la proporción de desocupados según el nivel educativo, comparando el primer trimestre de 2004 y 2024. En 2004, la mayoría de los desocupados no habían completado la educación primaria o secundaria. En cambio, en 2024, casi la mitad de los desocupados (41.86 %) tenía la secundaria completa. Es importante destacar que en 2024 la proporción de desocupados se concentra en niveles educativos más altos que en 2004. Un ejemplo claro es el aumento de desocupados con educación universitaria completa: en 2024, esta proporción se duplica en puntos porcentuales en comparación con 2004, pasando del 5.85 % al 12.79 %.
- La Tabla 2 muestra la proporción de desocupados según el rango de edad, comparando el primer trimestre de los años 2004 y 2024. En 2004, la mayoría de los desocupados tenía entre 20 y 30 años. Para 2024, la proporción de desocupados en este rango de edad aumentó en casi 15 puntos porcentuales. Además, se observa un incremento significativo en el grupo de 30 a 40 años, donde la proporción de desocupados es tres veces mayor que en 2004. De manera similar, en el grupo de 40 a 50 años, la proporción de desocupados se duplicó en comparación con 2004.

Tabla 1: Proporción de Desocupados por Nivel Educativo.

<b>Año</b>	<b>Nivel Educativo</b>	<b>Desocupados (%)</b>
<b>2004</b>	Primaria Incompleta (incluye educación especial)	8.77
	Primaria Completa	22.22
	Secundaria Incompleta	20.47
	Secundaria Completa	22.22
	Superior Universitaria Incompleta	20.47
	Superior Universitaria Completa	5.85
	Sin instrucción	0.00
<b>2024</b>	Primaria Incompleta (incluye educación especial)	9.30
	Primaria Completa	9.30
	Secundaria Incompleta	16.28
	Secundaria Completa	41.86
	Superior Universitaria Incompleta	10.47
	Superior Universitaria Completa	12.79
	Sin instrucción	0.00

Tabla 2: Proporción de Desocupados por Rango de Edad.

<b>Rango de Edad</b>	<b>2004 (%)</b>	<b>2024 (%)</b>
0-10	0.00	0.00
10-20	8.54	6.54
20-30	25.28	38.32
30-40	5.24	15.89
40-50	6.56	13.08
50-60	6.89	4.67
60-70	2.95	0.93
70-80	0.66	0.93
80-90	0.00	0.00
90-100	0.00	0.00
100-110	0.00	0.00

## Parte II: Clasificación

El objetivo de esta sección es intentar predecir si una persona está desocupada o no, a partir de ciertas variables de características individuales. Para eso, se realizará el siguiente procedimiento. Primero, se preparará la información en las submuestras de entrenamiento y prueba. Segundo se implementarán los métodos de Regresión Logística, Análisis Discriminante Lineal, K-Vecinos más Cercanos (KNN, por sus siglas en inglés) y *Naive Bayes* y se reportarán diferentes métricas de evaluación. Tercero, se compararán los resultados obtenidos con cada método para cada año y se analizarán las capacidades de predicción. Por último, se realizará la misma predicción para el subconjunto de la población que no respondió la encuesta.

### Inciso 1

Como primer paso se filtró la base de datos para cada año. Luego, para realizar el ejercicio de predicción en la submuestra de cada año se consideró como variable explicada

y a *Desocupado* y como variables explicativas al mismo vector de características individuales utilizadas en el inciso 2.d. (con excepción de *cat\_inac* e *IPCF*). Finalmente, se partió cada una muestra de prueba (*test*) y una de entrenamiento (*train*) conteniendo el 30 % y el 70 % de los datos, respectivamente. Ver *Python Script* para mayor detalle.

## Inciso 2

Una vez ordenada la información se procedió con la estimación de cuatro modelos de predicción: Regresión Logística, Análisis Discriminante Lineal (LDA), K-Vecinos más Cercanos (KNN con  $k=3$ ) y *Naive Bayes*.

En primer lugar, la Regresión Logística estima la probabilidad de que un individuo pertenezca a la categoría de desocupado, asumiendo que la relación entre las variables independientes y la dependiente es lineal. Por su parte, la LDA ajusta una densidad gaussiana a cada clase, asumiendo normalidad y una matriz de covarianza común entre todas las clases. Además, busca una combinación lineal de características  $X$  que mejor identifique las diferencias entre clases. A su vez, el algoritmo de clasificación no paramétrico de K-Vecinos más Cercanos se basa en la distancia entre puntos en la muestra de entrenamiento y no en una combinación lineal de características. KNN clasifica a un individuo como ocupado o desocupado según la clase más frecuente de los  $k$  (en este caso 3) vecinos más cercanos, en términos de distancia euclidiana. Por último, el método de *Naive Bayes* también calcula las probabilidades de que un individuo esté desocupado pero suponiendo que el conjunto de características individuales consideradas son independientes entre sí. Sin embargo, este supuesto es muy poco creíble en este ejemplo, donde tenemos fuertes correlaciones entre las  $x$ .

Para evaluar el desempeño predictivo de cada modelo, se presentan las métricas de *Accuracy* (exactitud), AUC y las matrices de confusión de cada modelo para cada año. La medida de *Accuracy* mide la proporción de predicciones correctas sobre el total de predicciones. A pesar de ser una medida simple y clara, si la proporción de desocupados y ocupados no está balanceada, la exactitud puede ser engañosa y poco ilustrativa del ejercicio de predicción. La medida de AUC se refiere al área debajo de la Curva ROC de los *scores* predichos. La Curva ROC muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para todos los umbrales del modelo. Por lo que, al medir el área por debajo, muestra qué tan bien el modelo predice como desocupados a los verdaderos desocupados: a mayor AUC, mayor precisión predictiva. Por último, se muestran las matrices de confusión correspondientes. Éstas resumen el rendimiento del clasificador mostrando el número de predicciones correctas e incorrectas para cada categoría. En nuestro ejemplo se presentan en las filas las  $Y$  reales y en las columnas las  $Y$  predichas; la intuición de las matrices es la siguiente:

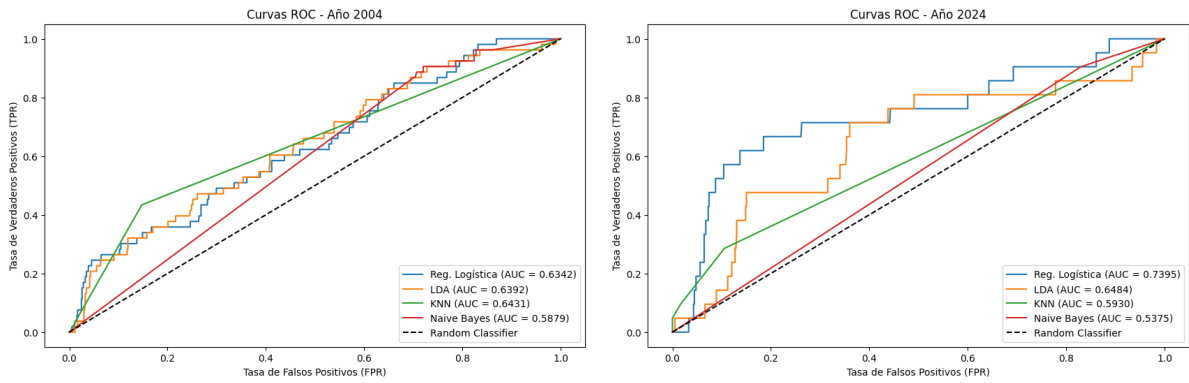
$$\begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix}$$

En la tabla 3 se presentan las medidas de precisión de *Accuracy* (exactitud), AUC y las matrices de confusión para cada modelo y cada año. También se muestran las curvas ROC en la figura 5.

Tabla 3: Métricas de Performance y Matrices de Confusión por Año y Modelo.

Año	Métrica	Log. Reg.	LDA	KNN (k=3)	Naive Bayes
2004	Accuracy	0.9321	0.9248	0.9212	0.2352
	AUC	0.6342	0.6392	0.6431	0.5879
	M. Confusión	$\begin{bmatrix} 769 & 3 \\ 53 & 0 \end{bmatrix}$	$\begin{bmatrix} 769 & 3 \\ 53 & 0 \end{bmatrix}$	$\begin{bmatrix} 757 & 15 \\ 50 & 3 \end{bmatrix}$	$\begin{bmatrix} 145 & 627 \\ 4 & 49 \end{bmatrix}$
2024	Accuracy	0.9489	0.9617	0.9537	0.1949
	AUC	0.7395	0.6484	0.5930	0.5375
	M. Confusión	$\begin{bmatrix} 594 & 11 \\ 21 & 0 \end{bmatrix}$	$\begin{bmatrix} 602 & 3 \\ 21 & 0 \end{bmatrix}$	$\begin{bmatrix} 595 & 10 \\ 19 & 2 \end{bmatrix}$	$\begin{bmatrix} 103 & 502 \\ 2 & 19 \end{bmatrix}$

Figura 5: Curvas ROC para el 2004 y 2024.



### Inciso 3

En base a los resultados presentados en el inciso anterior, los métodos que mejor predicen tanto en 2004 como en 2024 son los modelos de vecinos más cercanos (KNN) y el de regresión logística, respectivamente. Es indispensable destacar, antes que nada, que para arribar a esta conclusión hemos asumido simetría en la función de pérdida frente a los dos tipos de errores posibles: predecir como desocupado a a un ocupado o viceversa.

Para el año 2004, el modelo KNN se destaca por su desempeño equilibrado en las métricas de *accuracy* (0.9212) y AUC (0.6431). Aunque no tiene la mayor exactitud (siendo superado ligeramente por la regresión logística y LDA), es el modelo que mayor AUC presenta, lo cual indica una mayor capacidad para discriminar entre desocupados y ocupados en diferentes umbrales. En este caso, dado que la cantidad de ocupados y desocupados no está para nada balanceada, le asignamos una mayor al AUC.

En este sentido, el modelo de regresión logística obtiene un *accuracy* levemente superior (0.9321) pero con un AUC más bajo (0.6342), sugiriendo que puede ser menos robusto al variar los umbrales de decisión. A pesar de esto, sigue siendo una alternativa competitiva. En contraste, el modelo de Naive Bayes presenta un desempeño considerablemente inferior en ambas métricas (*accuracy* de 0.2352 y AUC de 0.5879), por lo que no se considera adecuado para este año.

Para el año 2024, el modelo de regresión logística sobresale como el más confiable. Su *accuracy* de 0.9489 es muy superior a la de los demás métodos, y su AUC de 0.7395 es también la más alta del grupo. El modelo de LDA también muestra un desempeño competitivo, con el mayor Accuracy (0.9617), aunque su AUC (0.6484) es inferior al de la regresión logística, lo que sugiere que podría ser menos robusto para discriminar entre

desocupados y ocupados en ciertos umbrales. Por otro lado, el modelo de KNN y el de *Naive Bayes* presentan un desempeño más limitado, con AUC y *accuracy* inferiores respecto a los dos modelos anteriores.

El supuesto de simetría frente a los dos tipos de errores posibles nos ha permitido centrar nuestro análisis en las medidas de AUC y *accuracy*. Sin embargo, en muchos contextos (por ejemplo, políticas públicas), esta simetría no es realista ya que un error puede ser más grave que el otro. De ser así, deberíamos enfocar un nuestro análisis en las matrices de confusión, donde podemos observar la probabilidad de cometer un tipo de error u el otro.

## **Inciso 4**

Dado que la única persona que no respondió acerca de su condición de actividad corresponde a la base de datos del año 2004, hemos utilizado el modelo de vecinos más cercanos (con  $k = 3$ ) para predecir si está persona es desocupada o no. El modelo identifica que está persona no es desocupada.