Trabajo práctico N° 3

Informe escrito

Autoras

Clara Bernhardt

María Teresa Laszeski

Rosario Luque

Profesores

María Noelia Castillo

Ignacio Spiousas

Asignatura

Ciencia de Datos

Semestre y año de presentación

2° semestre 2024

Parte I: Analizando la base

1. Identificación de personas desocupadas

La Encuesta Permanente de Hogares (EPH) identifica a las personas desocupadas mediante criterios específicos aplicados durante una semana de referencia. Se considera desocupada a una persona que cumple tres condiciones simultáneamente: no tiene ocupación (ni siquiera una hora de trabajo), está buscando trabajo activamente y está disponible para comenzar a trabajar. Esta información se recolecta a través de encuestas a hogares en aglomerados urbanos seleccionados. La tasa de desocupación se calcula como el porcentaje de la población desocupada respecto a la población económicamente activa. Este método se enfoca en la desocupación abierta, excluyendo otras formas de precariedad laboral, como el subempleo, que se miden con indicadores separados (INDEC, 2024).

2. Preprocesamiento de datos

En el proceso de limpieza y preparación de datos, se abordó la tarea de unificar y depurar las bases de la Encuesta Permanente de Hogares (EPH) correspondientes al primer trimestre de 2004 y 2024. Inicialmente, se descargaron los archivos en formatos .dta y .xls, respectivamente, desde la página del INDEC, para luego cargarlos en Visual Studio Code. Con el fin de facilitar la convergencia entre ambas bases, se procedió a convertir todos los nombres de variables a minúsculas. Posteriormente, se aplicó un filtro para retener únicamente los datos correspondientes a los aglomerados de Ciudad Autónoma de Buenos Aires (32) y Gran Buenos Aires (33), seguido de la concatenación de ambas bases en un único dataframe. En este proceso, se tomó la decisión de retener sólo las columnas de interés para el análisis posterior, específicamente, ano4, ch04, ch06, ch07, ch08, nivel_ed, estado, cat_inac e ipcf.

A continuación, se procedió a eliminar los valores negativos de las variables de interés. La variable ch06 es la correspondiente a edad y se observó que tenía 182 valores negativos, los cuales fueron eliminados. A su vez, observamos que ninguna columna tenía valores faltantes. Se hipotetiza que esto es producto de la función utilizada para leer la base de datos del 2004, la cual transforma las celdas vacías en ceros (0).

3. Gráficos y visualizaciones

Para el análisis de la composición por sexo, se crearon dataframes específicos para cada año (df_2004 y df_2024), verificando con la condición *if* la ausencia de valores vacíos que alertaría si no hubiera datos disponibles para alguno de los años. Utilizando las funciones value_counts() y sort_index(), se contabilizó la cantidad de hombres y mujeres en cada año. Para una presentación más clara de estos resultados, se creó un nuevo dataframe denominado "sexo_df" que contenía los conteos para ambos años, con los índices renombrados para mayor claridad (0 como "Varón" y 1 como "Mujer"). La visualización de estos datos se realizó mediante un gráfico de barras utilizando la biblioteca matplotlib.pyplot, lo que permitió una comparación visual inmediata de la composición por sexo entre 2004 y 2024. En la **Figura 1** se puede observar que en 2004 hay mayor composición de varones y mujeres en comparación con el 2024. En ambos años, las mujeres superan a los varones en cantidad, siendo en 2004 una cantidad de 3988 mujeres reportadas y en el 2024 una cantidad de 3651 mujeres reportadas. Por otro lado, la composición de hombres en 2004 es mayor a la composición del 2024, siendo en 2004 un número de 3528 hombres reportados y en 2024 una composición de 3349 hombres reportados.

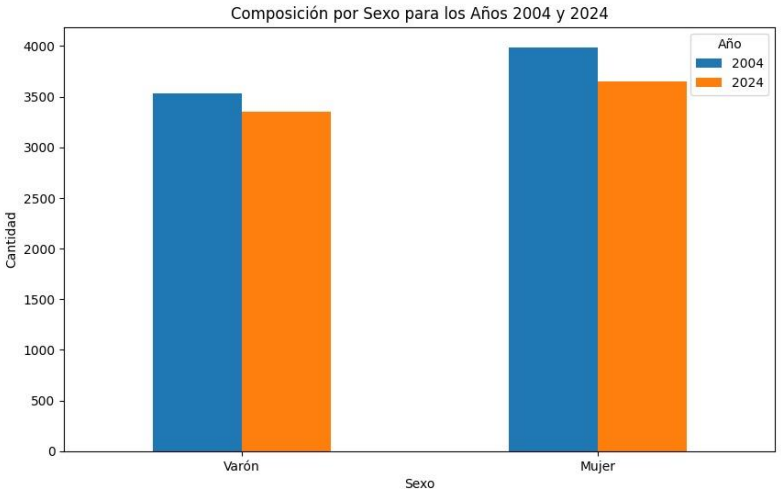

Figura 1. Histograma de la composición por sexo en los años 2004 y 2024.

Para realizar las matrices de correlación, primero renombramos algunas variables: "ch04" por "sexo", "ch06" por "edad", "ch07" por "estado_civil", "ch08" por "cobertura_medica". Luego, se transformaron en dummies las variables categóricas, esto es, reestructuramos las variables de “sexo”, “estado civil”, “cobertura médica”, “nivel ed”, “estado” y “cat_inac”. Se decidió realizar este paso con el fin de que la matriz de correlación fuera interpretable. A su vez, se partió la base de datos por año, de forma tal de tener un dataframe para el 2004 (df_2004) y otro para el 2024 (df_2024). Las matrices resultantes se visualizaron utilizando herramientas de representación gráfica, lo que permitió una interpretación intuitiva de las relaciones entre variables para cada año.

En las matrices de correlación de 2004 (**Figura 2**) y 2024 (**Figura 3**), se observan algunos cambios en las relaciones entre variables a lo largo de dos décadas. La correlación entre edad y estado civil se ha debilitado ligeramente, pasando de -0.48 en 2004 a -0.36 en 2024, lo que podría indicar una evolución en los patrones matrimoniales. La relación entre edad y nivel educativo se ha fortalecido, aumentando de 0.23 a 0.31, sugiriendo un incremento en la educación continua o acceso a la educación superior en edades más avanzadas. La cobertura médica muestra una correlación más fuerte con la edad en 2024 (0.29) en comparación con 2004 (0.22), posiblemente reflejando mejoras en el sistema de salud o mayor conciencia sobre la importancia de la cobertura médica con la edad. Es notable que la variable "estado_2" ha pasado de tener correlaciones casi nulas en 2004 a mostrar una relación más significativa con el estado civil (-0.33) en 2024, lo que podría indicar cambios en la dinámica entre situación laboral y estado civil. Las correlaciones con el ingreso per cápita familiar (ipcf) se mantienen relativamente estables, aunque con ligeros cambios. En general, mientras algunas relaciones se han fortalecido, otras se han debilitado, reflejando posibles cambios socioeconómicos y demográficos en la sociedad durante este período de 20 años.

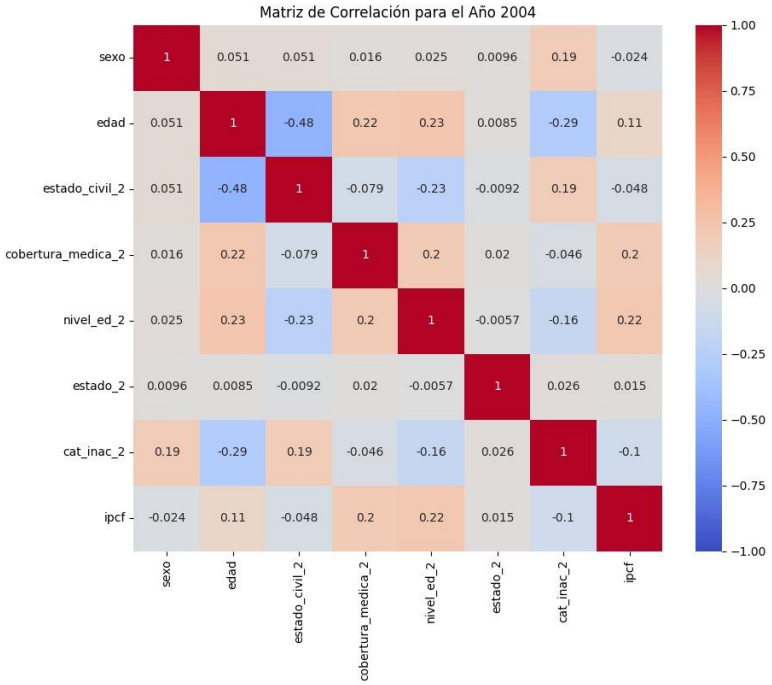

Figura 2. Matriz de correlaciones de las variables sexo, edad, estado_civil_2, cobertura_medica_2, nivel_ed_2, estado_2, cat_inac_2 e ipcf para el año 2004.

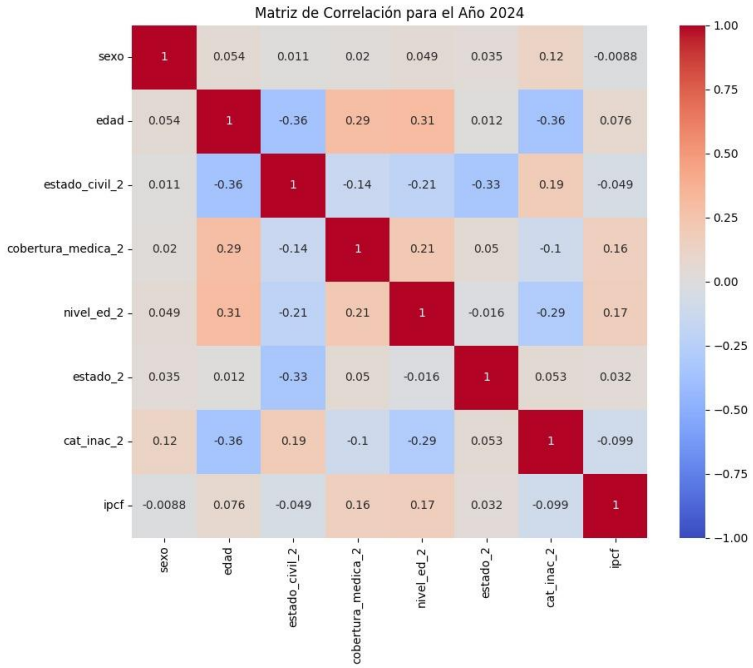

Figura 3. Matriz de correlaciones de las variables sexo, edad, estado_civil_2, cobertura_medica_2, nivel_ed_2, estado_2, cat_inac_2 e ipcf para el año 2024.

Finalmente, para el análisis del ingreso per cápita familiar (IPCF) en relación con el estado laboral, se aprovecharon nuevamente los dataframes df_2004 y df_2024. La variable "estado" se utilizó como base para categorizar a los individuos en tres grupos: Ocupados (1), Desocupados (2) e Inactivos (3). En la base de datos podemos observar que hay 6303 “Ocupados”, 839 “Desocupados” y 5462 “Inactivos”.

Se calculó la media del IPCF, siendo este el monto del ingreso per cápita familiar, para cada una de estas categorías en ambos años. Los resultados se consolidaron en un nuevo dataframe que permite una visualización clara de las medias de IPCF por estado laboral y año.

Al comparar los ingresos entre 2004 y 2024, se realizó un ajuste por inflación utilizando un factor de 436,16, calculado a partir de la variación en los ingresos de los ocupados durante ese período. Este factor se aplicó a los valores de 2004 para obtener cifras comparables en término de poder adquisitivo de 2024.

Los ocupados mantuvieron su poder adquisitivo, con un ingreso que se ajustó de 476,07 en 2004 a 207.644,85 en 2024, reflejando exactamente la tasa de inflación del período. Los inactivos experimentaron una ligera disminución en su poder adquisitivo real, pasando de un equivalente ajustado de 137.775,55 (315,89 * 436,16) a 130.704,61, indicando una pequeña pérdida en términos reales. Sin embargo, el grupo más afectado fue el de los desocupados, cuyo ingreso ajustado por inflación debería haber alcanzado 97.801,35 (224,23 * 436,16), pero en realidad solo llegó a 85.019,14, evidenciando una pérdida significativa de ingreso per cápita. Estos datos, ajustados mediante el factor de inflación calculado, sugieren que mientras la situación económica de los ocupados se mantuvo estable en términos reales, los inactivos experimentaron un ligero retroceso, y los desocupados enfrentaron un deterioro más pronunciado en su capacidad económica relativa durante estas dos décadas.

Luego, se hizo un análisis de las respuestas sobre la condición de actividad en la EPH. El objetivo fue identificar cuántas personas en la EPH no respondieron la pregunta sobre su condición de actividad. Esto es relevante debido al creciente problema de falta de datos en las encuestas, lo cual afecta la representatividad y precisión de los resultados. La cantidad de personas que no respondieron su condición de actividad fue de 51, lo cual podría ser un problema para estudios laborales, de ingresos y de estructura demográfica. A continuación, se presentan algunas de las características de los individuos que no respondieron su condición de actividad (**Tabla 1**). La ausencia de respuestas puede deberse a varias razones, como dificultades en la encuesta, falta de voluntad de los encuestados para responder ciertas preguntas o errores de registro. La falta de respuesta en variables clave como la condición de actividad podría inducir sesgos si los datos faltantes no son debidos al azar, afectando la representatividad de la muestra en análisis posteriores.

Cantidad de personas que no respondieron su condición de actividad: 51

	ano4	sexo	edad	estado_civil	cobertura_medica	nivel_ed	estado	cat_inac	ipcf	estado_civil_2	cobertura_medica_2	nivel_ed_2	estado_2	cat_inac_2
1592	2004.0	0.0	21.0	5.0	3.0	3.0	0.0	0.0	40.0	1.0	1.0	3.0	0.0	0.0
2208	2004.0	1.0	49.0	4.0	4.0	2.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0
2209	2004.0	0.0	24.0	5.0	1.0	3.0	0.0	0.0	0.0	1.0	1.0	3.0	0.0	0.0
2210	2004.0	1.0	20.0	5.0	4.0	4.0	0.0	0.0	0.0	1.0	0.0	4.0	0.0	0.0
2479	2004.0	1.0	35.0	3.0	4.0	5.0	0.0	0.0	294.0	1.0	0.0	5.0	0.0	0.0

Tabla 1. Subconjunto de registros de personas de la Encuesta Permanente de Hogares (EPH).

A continuación, se calculó una composición de la Población Económicamente Activa (PEA) para los años 2004 y 2024, para observar cómo se distribuye esta variable y compararla entre los años (**Figura 4**). La PEA incluye a las personas que se encuentran ocupadas o desocupadas, es decir, aquellas que están activamente participando en el mercado laboral. El gráfico muestra la cantidad de personas que forman parte de la PEA y las que no. Se puede ver que, en 2004, la cantidad de personas en la No PEA es mayor que las PEA. En 2024 el patrón se invierte, y ahora la población PEA es ligeramente mayor a la No PEA se mantiene similar. En consecuencia, se puede ver que los patrones son inversos, que podría mostrar transición en la composición demográfica o en el comportamiento laboral. Se podrían analizar factores como el envejecimiento poblacional, la educación, la política laboral y otros aspectos socioeconómicos que podrían haber impulsado esta inversión de patrones entre 2004 y 2024, que sería fundamental para entender si los cambios observados reflejan tendencias estructurales o circunstancias específicas de cada año.

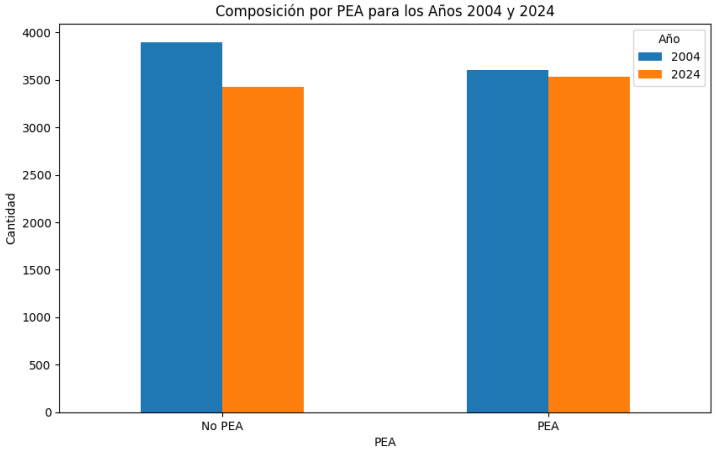

Figura 4. Composición por PEA para los años 2004 y 2024.

Después, se llevó a cabo una composición de la Población en Edad para Trabajar (PET: personas entre 15 y 65 años) y la comparación con la PEA en 2004 y 2024, para observar diferencias entre quienes están en edad de trabajar y quienes efectivamente participan en el mercado laboral. El gráfico muestra la cantidad de personas en edad de trabajar (PET) y fuera de esa categoría (No PET) en los años 2004 y 2024 (**Figura 5**). Además, la tabla comparativa entre PEA y PET permite observar algunas tendencias (**Tabla 2**). En primer lugar, en ambos años, la mayoría de las personas están en la categoría de PET, reflejando que gran parte de la población está en edad de trabajar. También, hay una disminución en la cantidad de personas tanto en la categoría No PET como en la PET en 2024 en comparación con 2004. Esto podría deberse a cambios demográficos o variaciones en el tamaño de la muestra. En segundo lugar, comparando la PEA con la PET, se puede ver que, aunque gran parte de la PET coincide con la PEA, existen diferencias; no todas las personas en edad de trabajar participan activamente en el mercado laboral (PEA). Además, en 2004, hay 3899 personas fuera de la PEA, aunque 3424 están en edad de trabajar. Esto sugiere que algunas personas en edad de trabajar no están económicamente activas. Por último, en 2024, también se observa que hay personas en la PET (3607) que no pertenecen a la PEA, lo que indica que no todas las personas en edad de trabajar están económicamente activas.

Finalmente, se hizo un análisis de la desocupación en 2004 y 2024, con el objetivo de determinar si había algún cambio en la tasa de desocupación en el tiempo. La desocupación se define en este contexto como aquellas personas que están en la Población Económicamente Activa (PEA) pero no tienen empleo. Los resultados muestran una disminución en la cantidad de personas desocupadas entre los dos años analizados, 528 para el 2004 y 311 para el 2024. La reducción en la cantidad de personas desocupadas en 2024 en comparación con 2004 podría indicar una mejora en las condiciones del mercado laboral o un cambio en la estructura de la población económicamente activa (PEA) en el tiempo. Sin embargo, estos resultados también podrían estar influenciados por factores externos, como cambios en las políticas de empleo, variaciones económicas, o diferencias en la composición de la muestra de cada año.

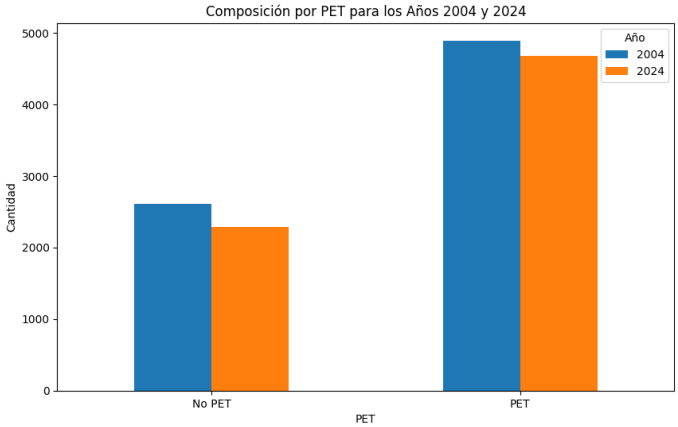

Figura 5. Composición por PET para los años 2004 y 2024.

	PEA 2004	PEA 2024	PET 2004	PET 2024
0	3899	3424	2613	2284
1	3607	3535	4893	4675

Tabla 2. Tabla comparativa entre las categorías de PEA y PET en 2004 y 2024 para analizar las diferencias.

Parte II: Clasificación

Una vez realizado el preprocesamiento de datos y agregadas las columnas indicadas, se procedió a entrenar y evaluar distintos modelos con el objetivo de intentar predecir si una persona está desocupada o no. Para ello, se filtró la base de datos “respondieron” para solo retener las variables de año, sexo, edad, PEA, PET y desocupado, sumado a las variables de estado civil, cobertura médica, nivel educativo, estado y categoría de inactividad que fueron transformadas a dummys.

A continuación, se realizaron dos divisiones. En primer lugar, se partió la base de datos “respondieron” por año, quedando separados los datos del 2004 y de 2024. En segundo lugar, se dividió cada base de datos en una base de prueba (*test*) y en una de entrenamiento (*train*) utilizando la función *train_test_split*. Cada base de entrenamiento comprendía el 70% de los datos, y la semilla utilizada fue 101. Una vez realizadas las dos divisiones, se estableció a “desocupado” como la variable dependiente en la base de entrenamiento (vector y) por lo que el resto de las variables, menos la columna de año, fueron las variables independientes (matriz X). Además, se agregó una columna de unos (1) a la matriz de variables independientes.

Luego se procedió a entrenar y testear cuatro modelos para cada año: regresión logística, análisis discriminante lineal (LDA), KNN con k=3 y Naive Bayes. A su vez, para evaluar la *performance* de los distintos modelos, se reportaron la matriz de confusión, la curva ROC, los valores de AUC y de Accuracy de cada uno para cada año.

Para el año 2004, los resultados encontrados reflejan que el mejor modelo para predecir la variable "desocupados" es el modelo de regresión logística. Por un lado, la regresión tiene un valor de *accuracy* de 0.9325, que es el más alto de los cuatro modelos. Por el otro, el área bajo la curva ROC (AUC) es de 0.8779, también es el más alto. La **Tabla 3** permite visualizar la comparación de estos dos parámetros. Sin embargo, el gráfico de la curva ROC refleja que LDA y Naive Bayes tienen apenas peor capacidad para distinguir entre clases (desocupados y no desocupados) en el año 2004 (**Figura 6**).

	Regresión Logística	LDA	KNN	Naive Bayes
Accuracy	0.9325	0.9312	0.9205	0.6394
AUC	0.8779	0.8717	0.6304	0.8484

Tabla 3. Valores de accuracy y AUC para la regresión logística, LDA, KNN con k=3 y Naive Bayes para la predicción de desocupados del año 2004.

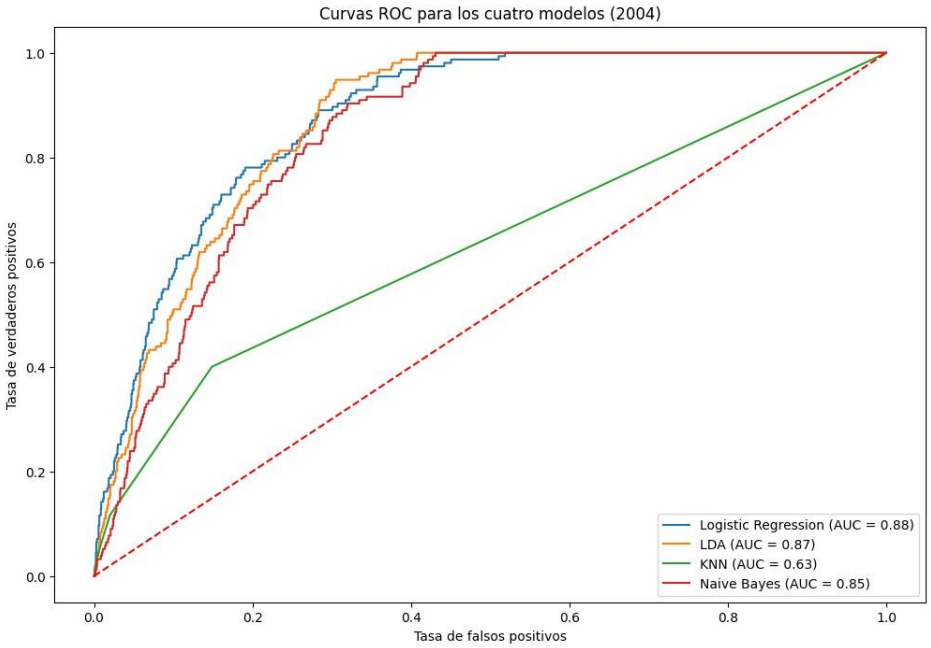

Figura 6. Curva ROC para la regresión logística, LDA, KNN con k=3 y Naive Bayes para la predicción de desocupados en el año 2004. La línea punteada roja representa la línea de azar.

Por su parte, para el año 2024, los resultados muestran que el mejor modelo para predecir los "desocupados" es LDA. Por un lado, el valor de *accuracy* para LDA es de 0.9521, igual al de regresión logística pero mayor a KNN y Naive Bayes. Por el otro, el AUC es de 0.8709, superior a los valores de los otros tres modelos. La **Tabla 4** permite visualizar la comparación de estos dos parámetros. El gráfico de la curva ROC refleja que LDA tiene gran capacidad para distinguir entre clases (desocupados y no desocupados) en comparación con los otros tres modelos para el año 2024 (**Figura 7**).

	Regresión Logística	LDA	KNN	Naive Bayes
Accuracy	0.9521	0.9521	0.9444	0.9521
AUC	0.8141	0.8709	0.6003	0.7185

Tabla 4. Valores de accuracy y AUC para la regresión logística, LDA, KNN con k=3 y Naive Bayes para la predicción de desocupados del año 2024.

Finalmente, se realizó la predicción de las personas desocupadas dentro de la base “norespondieron”. En primer lugar, se agregó la variable PET y se realizó el mismo filtrado de variables que se hizo previamente para la base “respondieron”. Es importante entender que esta base no cuenta ni con la variable PEA ni con desocupado. Por lo tanto, para poder utilizar los modelos entrenados y testeados, la columna PEA fue rellenada con ceros (0) y la variable desocupado se creará a partir de la predicción. Nuevamente, para realizar el proceso se dividió la base por año, de forma tal de aplicar la regresión logística a los datos del 2004 y LDA a los del 2024. Los resultados demuestran que de las personas que no respondieron su estado, 13 de 51 fueron predichas como desocupadas, lo cual representa un 25.49% del total de personas que no reportaron su estado.

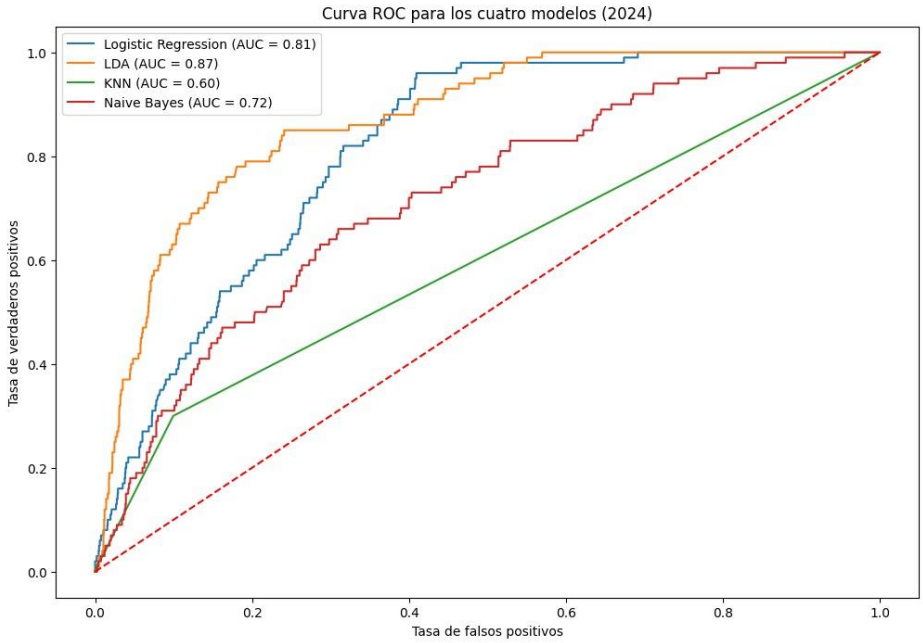

Figura 7. Curva ROC para la regresión logística, LDA, KNN con k=3 y Naive Bayes para la predicción de desocupados en el año 2024. La línea punteada roja representa la línea de azar.

Referencias

INDEC. (2024). *Mercado de trabajo. Tasas e indicadores socioeconómicos* (EPH) (Trabajo e Ingresos Vol. 8, N° 7). Recuperado de https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-31-58