Perfecto el trabajo!
Código: muy bueno!!

NOTA: 10

Machine Learning para Economistas
Trabajo Práctico III

Análisis Descriptivo y Predicción de Desocupación

Manuel Díaz de la Fuente, Diego Fernández Meijide y Sofía Kastika

Profesor: Walter Sosa Escudero
Asistente: Tomás Pacheco

Noviembre 2024

Parte I: Analizando la base

El objetivo de esta sección es el de la realización de una análisis descriptivo utilizando datos de la Encuesta Permanente de Hogares (EPH) de Bahía Blanca para los primeros trimestres de los años 2004 y 2024.

Inciso 1

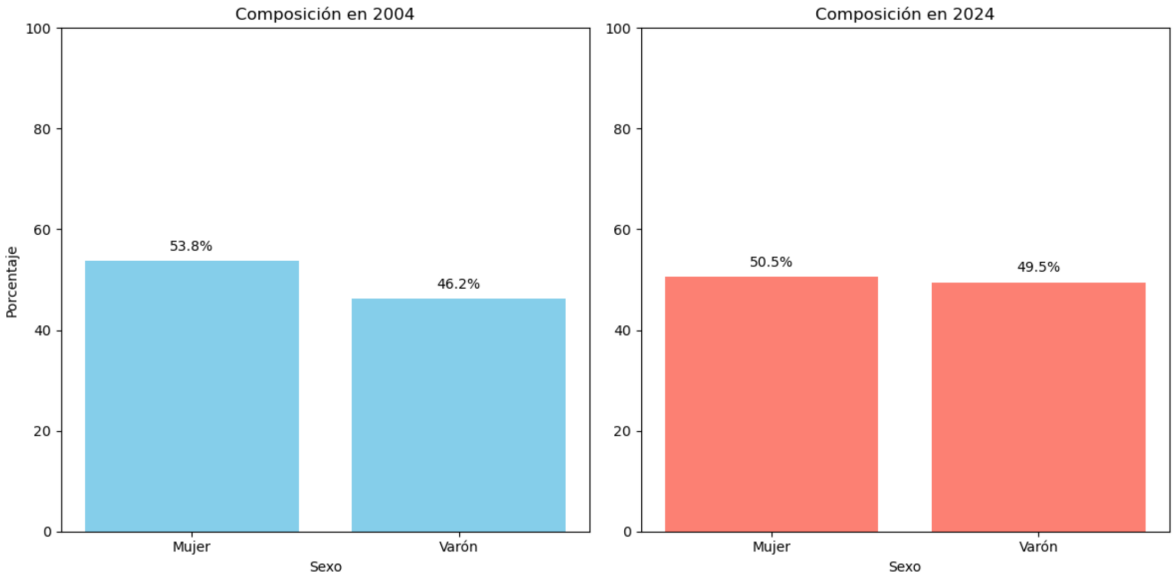
El Instituto Nacional de Estadística y Censos (INDEC) clasifica a las personas desocupadas como aquellas que no realizaron una actividad remunerada durante la semana anterior a la realización de la encuesta, que al mismo tiempo se encuentran en condiciones de trabajar y están dispuestas a hacerlo si la ocasión se presentara, y que se encuentran en búsqueda activa de empleo -es decir, que llevaron a cabo acciones para encontrar empleo durante la semana anterior a la encuesta-. Esta categoría no incluye al subconjunto de la población que actualmente no está buscando trabajo -población inactiva-, ni a la que se se encuentra subocupada -que trabaja menos de 35 horas semanales por causas involuntarias-.[1]

Inciso 2

Nuestro grupo trabajará con los datos de Bahía Blanca, cuyo código en la Encuesta Permanente de Hogares (EPH) es el 3. Por ello, luego de descargar las bases de datos de los primeros trimestres de 2004 y de 2024 desde la página web del INDEC, filtramos la información de esta ciudad[2]. Asimismo, también eliminamos de ambas bases de datos aquellas observaciones que presentan edades negativas. Las demás variables a utilizar en el resto del trabajo práctico presentan valores en rangos razonables.

En la Figura 1 se observa que en ambas muestras la composición de la población se divide de forma equitativa entre hombres y mujeres. Igualmente, notamos un leve aumento en el porcentaje de mujeres, pasando de representar un 53.8 % en 2004 a un 50.5 % en 2024.

Figura 1: Composición por sexo. 2004 y 2024



Nota: elaboración propia en base a datos de la Encuesta Permanente de Hogares. Las barras muestran la proporción de varones y mujeres en las muestras de 2004 y de 2024.

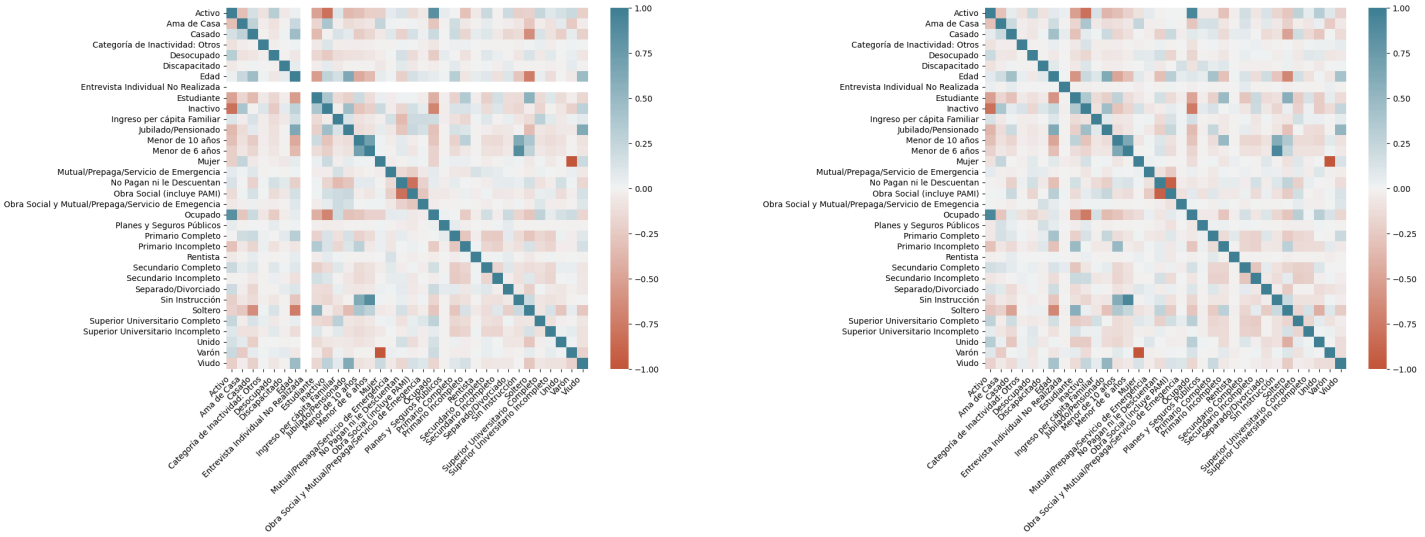
[1] Definición obtenida del Glosario del Instituto Nacional de Estadística y Censos: https://www.indec.gob.ar/indec/web/Institucional-Indec-Glosario

[2] Cabe aclarar que, si bien la EPH es una encuesta que, a diferencia del Censo Nacional de Población, Hogares y Vivienda, no incluye a la totalidad de los habitantes, consideramos a nuestras muestras representativas de la población de Bahía Blanca en 2004 y 2024 para sus interpretaciones

Las Figuras 2a y 2b muestran las matrices de correlación entre las variables relevantes de nuestras bases de datos. La mayor parte de las correlaciones parecen ser débiles. Ambas matrices sugieren la ausencia de fuertes correlaciones entre ser discapacitado y el resto de las variables, entre el ingreso per cápita familiar y el resto de las variables, entre contar con planes y seguros públicos y el resto de las variables, y entre ser rentista y el resto de las variables.

Algunas de las correlaciones fuertes resultan triviales, como por ejemplo que ser varón está fuertemente y negativamente correlacionado con ser mujer, o que estar activo está fuertemente y positivamente correlacionado con estar ocupado. Cabe resaltar que encontrarse desocupado no correlaciona fuertemente con ninguna variable.

Figura 2: Matrices de correlación para 2004 y 2024



(a) Matriz de correlación 2004

(b) Matriz de correlación 2024

Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares. Las celdas se encuentran coloreadas de acuerdo a la intensidad de la correlación entre cada una de las variables de los ejes.

Por otra parte, los Cuadros 1 y 2 indican que en nuestra muestra de Bahía Blanca de 2004 hay 91 personas desocupadas y 494 inactivas, y en la de 2024 38 personas desocupadas y 425 inactivas. Por ende, el total de desocupados en nuestra muestra total de individuos es de 129. La media del ingreso per cápita familiar en 2004 para los ocupados es de 399 pesos, para los desocupados es de 206 pesos y para los inactivos es de 315 pesos. En 2024 la media del ingreso per cápita familiar para los desocupados es de 105 mil pesos, para los ocupados es de 170 mil pesos y para los inactivos es de 117 mil pesos.

Cuadro 1: Cantidad de individuos y media de Ingreso Per Cápita Familiar (IPCF) por condición de actividad 2004

Categoria	Cantidad individuos	Media IPCF
Inactivos	494	315
Ocupados	440	399
Desocupados	91	206
Menores de 10 años	129	259

Nota: En la tabla se observan estadísticas descriptivas para el estado de ocupación y de Ingreso Per Cápita Familiar de la muestra de la EPH (INDEC) de Bahía Blanca para el año 2004.

Cuadro 2: Cantidad de individuos y media de Ingreso Per Cápita Familiar (IPCF) por condición de actividad 2024

Estado	Cantidad individuos	Media IPCF
Ocupados	432	169874
Inactivos	425	117397
Desocupados	38	105006
Menores de 10 años	102	96909
No responde	1	0

Nota: En la tabla se observan estadísticas descriptivas para el estado de ocupación y de Ingreso Per Cápita Familiar de la muestra de la EPH (INDEC) de Bahía Blanca para el año 2024.

Inciso 3

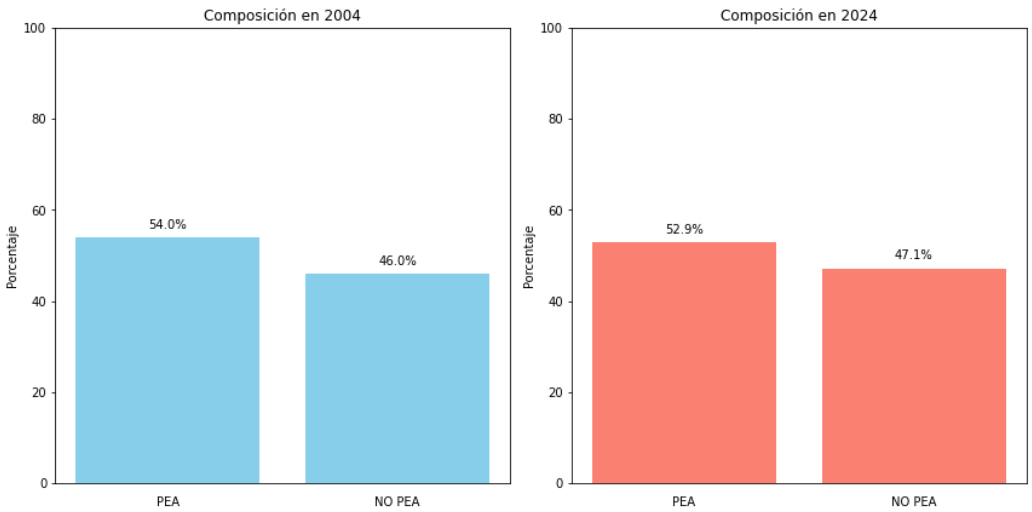
El Cuadro 2 muestra que solo una persona no respondió a la pregunta sobre la condición de actividad en Bahía Blanca en los años y trimestres analizados. Esta sola observación es del año 2024.

Inciso 4

Según el INDEC la Población Económicamente Activa (PEA) está compuesta por aquellas personas que tienen una ocupación o que buscan activamente una[3]. Es por ello que, en la Encuesta Permanente de Hogares, la PEA se identifica como la suma de los ocupados y desocupados. Para calcular el porcentaje de población económicamente activa, se divide la suma de ocupados y desocupados entre el total de individuos, que incluye también a los inactivos, menores de 10 años y aquellos que no respondieron sobre su condición de actividad.

En la Figura 3 puede observarse la composición de la PEA en Bahía Blanca para los años 2004 y 2024. En 2004, el porcentaje de población económicamente activa representaba un 54 % de la población, lo que equivale a casi la mitad de la población de Bahía Blanca. Este porcentaje no ha cambiado considerablemente en 20 años: en 2024, la población económicamente activa representa un 52.9 %, lo que implica una caída de 1.1 puntos porcentuales desde 2004 hasta la actualidad.

Figura 3: Composición por Población Económicamente Activa (PEA)



Nota: elaboración propia en base a datos de la Encuesta Permanente de Hogares. PEA hace referencia a la Población Económicamente Activa (personas que contestaron estar ocupadas o desocupadas en la encuesta) y NO PEA hace referencia a la población no económicamente activa (menores de 10 años, inactivos, personas que no respondieron la encuesta).

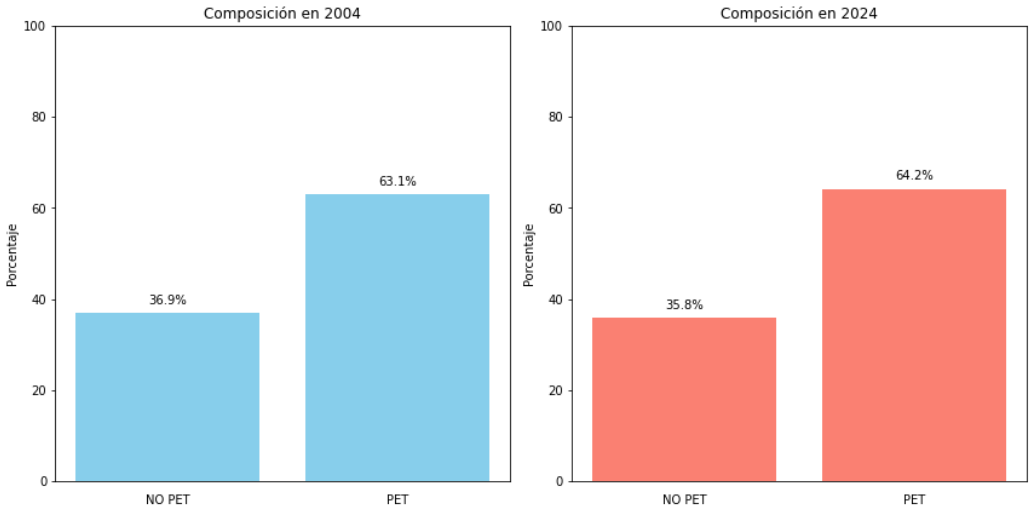
[3]Definición obtenida del Glosario del Instituto Nacional de Estadística y Censos: https://www.indec.gob.ar/indec/web/Institucional-Indec-Glosario

Inciso 5

Definimos a la Población en Edad para Trabajar (PET) como aquellas personas que tienen entre 15 y 65 años. En la Figura 4 puede observarse la composición por PET en Bahía Blanca para los años 2004 y 2024. Como puede observarse, en 2004 la población en edad para trabajar representaba un 63.1 %, mientras que en 2024 representa un 64.2 % lo cual implica que hubo un cambio muy pequeño en la población en edad para trabajar en 20 años. Específicamente, el cambio fue de 1.1 puntos porcentuales positivos.

Si se compara con lo obtenido en el inciso anterior, puede primeramente comentarse que la composición tanto por población económicamente activa como por edad para trabajar se mantuvo relativamente estable entre el 2004 y el 2024. Sin embargo, si bien observamos que en Bahía Blanca en la actualidad la proporción de población entre 15 y 65 años - es decir, en edad para trabajar- es 1.1 puntos porcentuales mayor que lo que era en el 2004, la población económicamente activa cayó en 1.1 puntos porcentuales en comparación con 2004.

Figura 4: Composición por Población en Edad para Trabajar (PET)



Nota: elaboración propia en base a datos de la Encuesta Permanente de Hogares. La PET hace referencia a la Población en Edad para Trabajar, es decir, a las personas entre 15 y 65 años. La NO PET hace referencia a la población que no está en edad para trabajar, es decir, los individuos menores de 15 años y mayores de 65.

Inciso 6

Con respecto a la desocupación, los datos presentados en los Cuadros 1 y 2 indican que, durante el primer trimestre de 2004, existían 91 personas desocupadas, cifra que disminuyó a 38 en el primer trimestre de 2024. Esta reducción puede explorarse más detalladamente a través de la proporción de desocupados por nivel educativo, cuyos resultados pueden observarse en el Cuadro 3.

Si analizamos la evolución por nivel educativo, observamos que la tasa de desocupación disminuyó en todos los niveles entre 2004 y 2024. Específicamente, para individuos con educación primaria completa, la proporción de desocupados se redujo de un 7.6 % en 2004 a un 2.5 % en 2024, lo que representa una disminución de 5.1 puntos porcentuales. Similarmente, aquellos con educación secundaria (completa e incompleta) mostraron reducciones significativas. Los desocupados con secundaria incompleta pasaron de representar un 10.3 % a un 4.5 %, una baja de 5,8 puntos porcentuales. En el caso de los que completaron la secundaria, la tasa de desocupados cayó de un 13.8 % a un 5.6 %, una reducción de 8.2 puntos porcentuales. En cuanto a la educación universitaria, los individuos con universitario incompleto vieron su tasa de desocupación decrecer de un 13,4 % en 2004 a un 8.5 % en 2024, una disminución de 4,9 puntos porcentuales. Aquellos con universitario completo presentaron una caída de un 3.5 % a un 1.3 % en sus tasas de desocupación, representando una caída de 2.2 pp.

Cuadro 3: Proporción de desocupados por nivel educativo en 2004 y 2024

Nivel Educativo	Proporción Desocupados 2004	Proporción Desocupados 2024
Primario Incompleto	0.020	-
Primario Completo	0.076	0.025
Secundario Incompleto	0.103	0.045
Secundario Completo	0.138	0.056
Universitario incompleto	0.134	0.085
Universitario completo	0.035	0.013

Notas: Elaboración propia en base a datos de la Encuesta Permanente de Hogares. No hay datos de desocupados con primario incompleto en 2024.

En conclusión, podríamos decir que la proporción de desocupados bajó en todos los niveles educativos de 2024 a 2004, pero en donde más se concentró el efecto fue en aquellas personas cuyo estudio máximo es secundario completo, con una caída de 8.2 pp.

Podríamos también analizar la proporción de desocupados por grupo etario. Como se observa en el Cuadro 4, la proporción de desocupados disminuyó en todos los grupos etarios entre 2004 y 2024. En el grupo de 10 a 19 años, esta proporción cayó de un 6.8 % en 2004 a un 2 % en 2024, lo que representa una disminución de 4.8 puntos porcentuales. Para el grupo de 20 a 29 años, la reducción es aún mayor: la proporción de desocupados descendió de 18.5 % a 9 %, reflejando una caída de 9.5 pp.

Cuadro 4: Proporción de desocupados por grupo de edad en 2004 y 2024

Edad	Proporción Desocupados 2004	Proporción Desocupados 2024
10-19	0.068	0.020
20-29	0.185	0.090
30-39	0.097	0.069
40-49	0.085	0.055
50-59	0.041	0.019
60-69	0.082	0.012
70-79	0.012	-

Notas: Elaboración propia en base a datos de la Encuesta Permanente de Hogares. No hay desocupados para los grupos etarios 0 a 9 y 80 a 89 para ambos años (lo cual tiene sentido dado que no son población en edad para trabajar) y para el grupo etario de 70 a 79 en 2024.

En los grupos de 30 a 39 y de 40 a 49 años, la proporción de desocupados disminuyó aproximadamente en 3 pp para cada grupo. También, se observó una reducción en el grupo de 50 a 59 años, con una baja de 2.2 pp. Finalmente, en el grupo de 60 a 69 años, la proporción de desocupados pasó de 8.2 % a 1.2 %, representando una reducción de 7 pp. En conclusión, la proporción de desocupados se redujo en todos los grupos etarios al comparar 2004 y 2024. Los grupos de 20 a 29 años y de 60 a 69 años fueron los que experimentaron las mayores disminuciones, con caídas de 9.5 pp y 7 pp, respectivamente.

Parte II: Clasificación

El objetivo de esta sección es predecir si una persona está desocupada o no, a partir de características individuales. Para ello, vamos a dividir a la muestra "respondieron" en una base de entrenamiento, con el 70 % de los datos, y una de prueba, con el 30 % de los datos. Luego, se estimaran distintos modelos utilizando la muestra de entrenamiento con el objetivo de predecir si los individuos de la muestra de prueba se encuentran desocupados. Una vez hecho esto,

utilizaremos el mejor modelo estimado para predecir si aquellas personas que no respondieron la pregunta acerca de su estado de ocupación se encuentran ocupadas o no. Presentamos a continuación las matrices de confusión de cada uno de los modelos estimados para los años 2004 y 2024. En estos resultados $\hat{Y} = 0$ indica que el modelo predice que la persona está ocupada, mientras que $\hat{Y} = 1$ indica una predicción de que la persona se encuentra desocupada.

Regresión Logística

Cuadro 5: Matriz de Confusión Regresión Logística

(a) 2004

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	312	0
$Y = 1$	32	0

(b) 2024

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	290	0
$Y = 1$	10	0

Como podemos observar, el modelo de regresión logística predice que la totalidad de los individuos se encuentran ocupados. Dado que la mayoría de las personas en la muestra de testeo se encuentran ocupadas, la tasa de verdaderos negativos es alta, pero los modelos no son capaces de identificar a las personas desocupadas a partir de sus características observables. Para este modelo tanto el *recall*[4] como la precisión[5] son de 0 %. corregir error latex

Análisis Discriminante Lineal

Cuadro 6: Matriz de Confusión Análisis Discriminante Lineal

(a) 2004

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	311	1
$Y = 1$	32	0

(b) 2024

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	289	1
$Y = 1$	10	0

Por otro lado, análisis discriminante presenta resultados muy similares a los de regresión logística, con la única diferencia de que para una observación realiza una errónea predicción de "desocupado". Al igual que con la regresión logística, tanto el *recall* como la precisión son de 0 %.

K-Nearest Neighbors

Cuadro 7: Matriz de Confusión KNN

(a) 2004

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	302	10
$Y = 1$	26	6

(b) 2024

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	286	4
$Y = 1$	10	0

KNN, a diferencia de regresión logística y análisis discriminante, predice "desocupados", pero lo hace a costa de una tasa de falsos positivos alta. Para este modelo en 2004, el recall es de 18 % y la precisión de 37.5 %. Ambos valores muestran un mejor desempeño de KNN para la predicción de la desocupación que los dos modelos anteriores en este

[4] Recall = Verdaderos Positivos/Positivos

[5] Precisión = Verdaderos Positivos/(Verdaderos Positivos + Falsos Positivos)

año. Para 2024 ambos valores son, nuevamente, 0 %. Es decir, en este año KNN muestra el mismo desempeño que los modelos anteriores.

Naive Bayes

Cuadro 8: Matriz de Confusión Naive Bayes

(a) 2004

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	143	169
$Y=1$	4	28

(b) 2024

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	290	0
$Y=1$	10	0

Por último, el clasificador Naive Bayes muestra un rendimiento muy diferente en ambos años. En 2004 parecería tender a predecir "desocupado" en una mayor proporción que "ocupado", por lo que presenta una tasa muy alta de falsos positivos. Sin embargo, muestra altos valores altos en recall -87.5 %- y en precisión -14.2 %-. En cambio, en 2024 se comporta igual que los modelos anteriores. La siguiente tabla muestra los *accuracy scores* de cada método de clasificación en ambos años:

Cuadro 9: Accuracy Score para los diferentes métodos de clasificación

Método de Clasificación	**2004**	**2024**
Regresión Logística	0.91	0.97
Análisis Discriminante Lineal	0.90	0.96
K- Nearest Neighbors	0.90	0.95
Naive Bayes	0.50	0.97

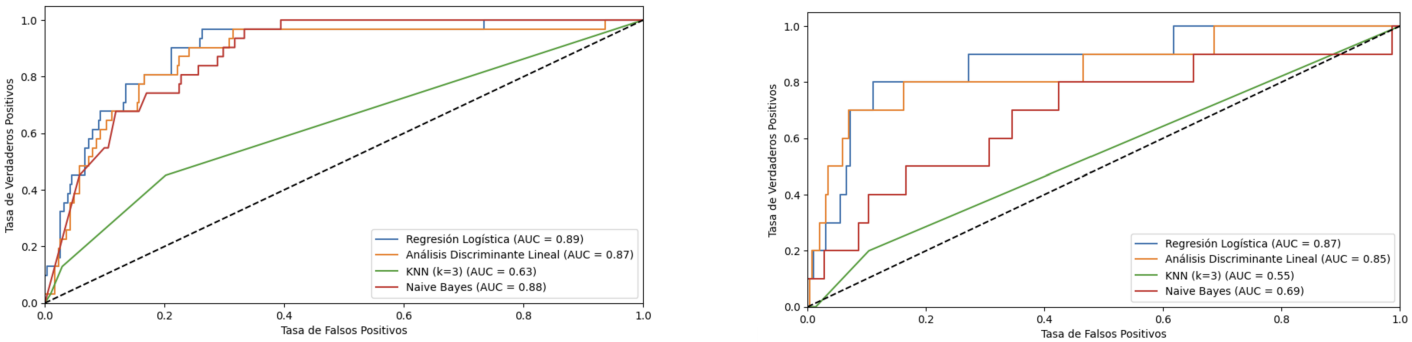
Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares.

En resumen, todos los métodos presentan un *score* relativamente alto, pero como vimos con las matrices de confusión, estos valores no cuentan toda la historia. Estos valores podrían explicarse por el hecho de que la gran mayoría de los individuos en la muestra de testeo se encuentran ocupados, por lo cual cualquier método que prediga "ocupado" con mayor frecuencia tendrá un *accuracy* elevado. Como contrapartida, tendrá muchos falsos negativos. A pesar de esto, la regresión logística es el clasificador con *accuracy* más alto en promedio para ambos años.

Curva ROC

Las curvas ROC muestran el trade-off entre la tasa de verdaderos positivos y la tasa de falsos positivos a medida que varía el umbral de decisión del clasificador. En otras palabras, representan la capacidad del modelo para identificar correctamente los verdaderos positivos frente a la probabilidad de cometer errores al clasificar instancias negativas como positivas (falsos positivos). En las siguientes figuras, se encuentran las curvas ROC para los distintos métodos de clasificación utilizados:

Figura 5: Curvas ROC para distintos métodos de clasificación en ambos años



(a) ROC 2004 **(b) ROC 2024**

Nota: Comparación de las curvas ROC entre los años 2004 y 2024. Elaboración propia.

Como se puede observar, las curvas ROC muestran que la Regresión Logísitca, el Naive Bayes Clasifier y el Análisis del Discriminante en 2004 presentan rendimientos muy similares, KNN presenta tasas de verdaderos positivos menores al resto de los clasificadores. Este "mal rendimiento" por parte de KNN podría explicarse porque es el clasificador que más predice "desocupado" y dado que la mayoría de los individuos en la muestra de testeo se encuentran ocupados, su tasa de falsos positivos es muy alta. En 2024, las diferencias entre clasificadores son un poco más grandes y, además, todos los clasificadores predicen peor. A pesar de esto, las observaciones hechas para 2004 también son válidas para 2024.

Para resumir la información de las curvas ROC en una única métrica, utilizamos la medida de *"area under the curve"* (AUC). En la siguiente tabla, podemos ver el AUC score para los diferentes clasificadores:

Cuadro 10: AUC Score de Diferentes Métodos de Clasificación en los Años 2004 y 2024

Método de Clasificación	**2004**	**2024**
Regresión Logística	0.80	0.74
Análisis Discriminante Lineal	0.83	0.77
K- Nearest Neighbors	0.65	0.55
Naive Bayes	0.78	0.69

Nota: Elaboración propia en base a datos de la Encuesta Permanente de Hogares.

Podemos observar que análisis discriminante lineal es el clasificador con el mayor AUC promedio. Por lo tanto, a pesar de que las diferencias en el *accuracy* de regresión logística y análisis discriminante favorecen a la regresión logística, estas diferencias son muy pequeñas y el análisis discriminante presenta un mejor rendimiento en cuanto al AUC. Por lo tanto, concluimos que el mejor modelo estimado es el de análisis discriminante. Por ello, lo utilizaremos para realizar las predicciones para aquellos individuos que no contestaron la pregunta acerca de su estado de ocupación.

Dado que en 2004 no hay ninguna persona que no haya respondido la pregunta acerca de su condición de actividad, no podemos realizar ninguna predicción utilizando el mejor modelo estimado en 2004. Sin embargo, dado que en 2024 hubo una única persona que no respondió, utilizando el modelo de análisis discriminante lineal estimado, predecimos que, dadas sus características[6], esa persona va a estar ocupada.

[6]Las características utilizadas para la predicción del modelo fueron: sexo, edad, estado civil, cobertura médica, nivel educativo e ingreso per cápita familiar