Ciencia de Datos
Trabajo Práctico № 4

Ignacio Spiousas
Noelia Romero

Ignacia Auguste
Trinidad Morán
Florencia Reverendo

Parte I: Análisis de la base de hogares y tipo de ocupación

1) Dos variables que creemos que pueden ser predictivas de la desocupación son CH09 (si la persona sabe leer y escribir) y CH10 (si la persona asiste o asistió a un establecimiento educativo). Esto se debe a que si una persona no sabe leer ni escribir es altamente probable que no trabaje, ya que hoy en día en la mayoría de los trabajos exigen estas dos habilidades. Además, si una persona asiste a un establecimiento educativo es posible que también esté desocupada, ya que la probabilidad de que sea menor de edad y, por consiguiente, esté desocupado, es más alta.

2) En este trabajo se utilizaron las bases de datos de la Encuesta Permanente de Hogares correspondientes al primer trimestre de 2004 y 2024. Cada año dispone de dos conjuntos de datos: uno sobre hogares y otro sobre personas. Para la elaboración de este análisis, estas bases se integraron utilizando los códigos de identificación de viviendas (CODUSU) y hogares (NRO_HOGAR), los cuales permitieron vincular los registros de hogares con las personas correspondientes. A partir de esta integración, se generó una base de datos conjunta que incluye únicamente información de la Ciudad Autónoma de Buenos Aires y el Gran Buenos Aires, resultando en un total de 14.698 registros.

3) Sobre esta base consolidada, se procedió a codificar las columnas CH04 (sexo), CH07 (estado civil), CH08 (cobertura médica), NIVEL_ED (nivel educativo), ESTADO (condición de actividad) y CAT_INAC (categoría de inactividad) como variables categóricas numéricas, asignando un valor numérico específico a cada categoría.

Con el objetivo de reducir posibles errores en los datos, se eliminaron aquellos registros con valores que carecían de sentido lógico. En particular, se descartaron los valores negativos en las variables CH06 (edad), PP03D (cantidad de ocupaciones), PP08D1 (ingresos mensuales), P21 (ingreso total habitual de la ocupación), IPCF (ingreso per cápita familiar), IX_TOT (cantidad de personas en el hogar) y CAT_INAC (categoría de inactividad), ya que estos valores deberían ser positivos para mantener la coherencia de los datos. Luego de la limpieza, la base de datos quedó conformada por 13.701 registros.

Además, se realizó un análisis de los datos faltantes en ciertas columnas de interés. Se identificaron los siguientes valores ausentes: 135 registros en la columna de edad (CH06), 13 registros en la columna de estado civil (CH07), 20 registros en la columna de cobertura médica (CH08), 1 registro en la columna de ingreso familiar per cápita (IPCF), 230 registros en la columna de cantidad de personas en el hogar (IX_TOT).

Dado que el número de datos faltantes era relativamente bajo en comparación con el tamaño total de la base, se optó por eliminar los registros con valores faltantes, excepto en el caso de la columna de edad (CH06). Consideramos que la edad es un dato fundamental para el análisis, por lo que decidimos estimar los valores faltantes utilizando la mediana de la edad correspondiente a cada grupo de nivel educativo. Este enfoque permitió preservar la relevancia de la variable sin introducir sesgos significativos en los datos.

4) Construcción de 4 variables:

PROP_OCUPADOS: Esta variable nos permite conocer la proporción de las personas que trabajan en un hogar.

PROP_MENORES: Nos muestra la proporción de menores de edad que habitan en el hogar.

DEP_ECO: El fin de esta variable es mostrar la proporción de habitantes activos vs inactivos de un hogar para ver de cuántas personas depende el hogar.

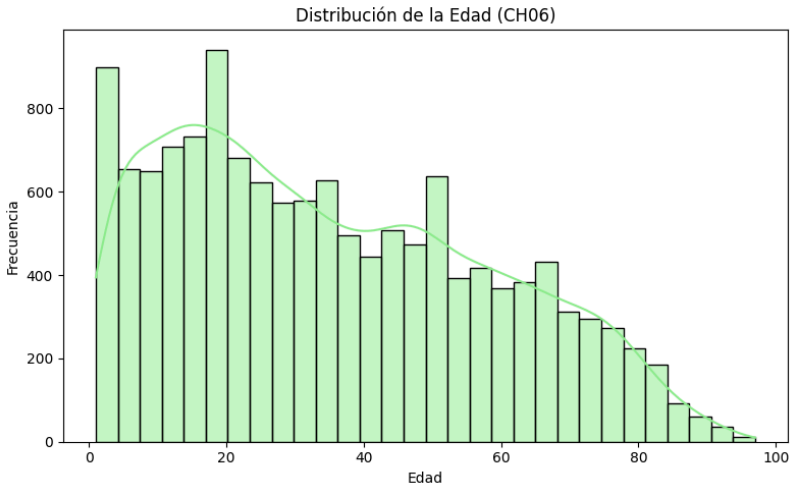
EDUC_PROM: Muestra el nivel educativo promedio de los habitantes del hogar.

DENSIDAD: Es la cantidad de personas que hay por habitación de la vivienda con el fin de saber cuántas personas hay por habitación.

5) En este apartado, se presentan las estadísticas descriptivas de tres variables de la (EPH) que consideramos relevantes para analizar la desocupación: edad (CH06), nivel educativo (NIVEL_ED) y categoría de inactividad (CAT_INAC). A continuación, se exponen los análisis y comentarios correspondientes a cada variable.

La edad es una variable clave para entender la desocupación, ya que los niveles de participación en el mercado laboral y las probabilidades de estar ocupado o desocupado varían significativamente según el grupo etario. Por ello, el análisis de esta variable resulta relevante para comprender la distribución de la desocupación en diferentes etapas de la vida.

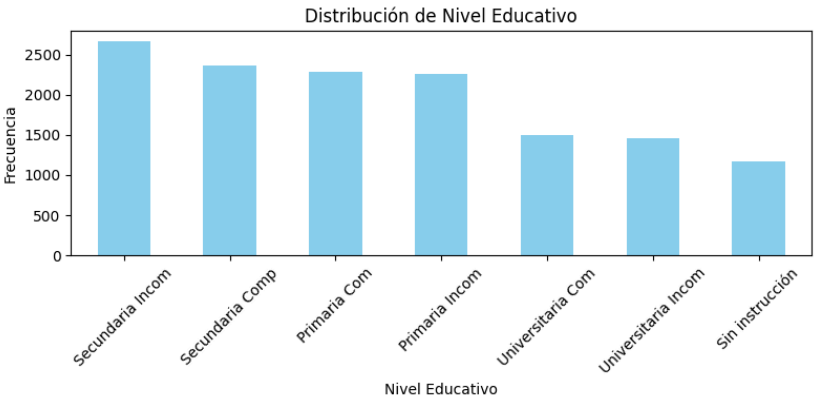
En cuanto a las estadísticas descriptivas obtenidas, la edad promedio de la población analizada es de 35.27 años, lo que indica que la mayoría de las personas están dentro de la edad laboral activa. La desviación estándar, de 23.14 años, evidencia una amplia dispersión en los valores de la muestra, lo que sugiere que la población incluye tanto a menores de edad como a adultos mayores. Los valores extremos de 1 y 97 años confirman la presencia de personas fuera de la edad laboral activa, como niños pequeños y jubilados. Al analizar los percentiles, se observa que el 25% de la población tiene 16 años o menos, lo que representa un segmento con baja participación en el mercado laboral. Por otro lado, la



mediana de 32 años refleja que la mitad de la población tiene esa edad o menos, mientras que el 75% alcanza hasta los 53 años, lo que abarca gran parte de la población económicamente activa.

El nivel educativo es un determinante clave en el análisis de la desocupación, ya que influye en la empleabilidad y en las oportunidades laborales de las personas. En general, niveles educativos más altos están asociados con una mayor empleabilidad, mientras que aquellos con menor educación enfrentan barreras significativas para acceder al mercado laboral. Por ello, esta variable resulta esencial para comprender las dinámicas de la desocupación.

En cuanto a la distribución de los niveles educativos, el grupo más frecuente es el de personas con educación secundaria incompleta, que representa el 19.42% de la población. Esto indica que una proporción considerable de individuos no alcanza a completar el nivel secundario, lo que puede limitar sus posibilidades de inserción laboral, especialmente en empleos que exigen una calificación mínima. Por su parte, el 17.29% de la población tiene secundaria completa, lo que señala que una cantidad significativa logra alcanzar este nivel, considerado un requisito básico en muchos ámbitos laborales.



El grupo de personas sin ningún tipo de instrucción constituye el 8.55% de la población, lo que sugiere que la mayoría cuenta con al menos algún grado de educación formal. Sin embargo, este segmento es particularmente vulnerable en el mercado laboral debido a la ausencia de calificaciones. En cuanto a los niveles superiores, el 10.95% de los individuos tiene educación universitaria incompleta, y el 10.66% ha completado este nivel, lo que en conjunto representa alrededor del 22% de la población. Esto evidencia que una parte importante accede a la educación superior, aunque no todos logran finalizarla.

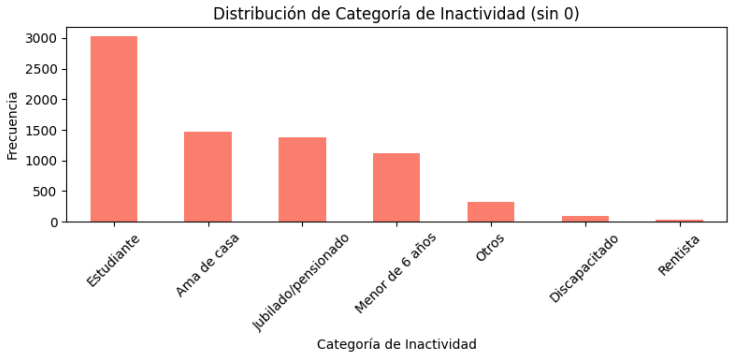
La variable CAT_INAC es esencial para identificar los individuos que no participan activamente en el mercado laboral, tales como jubilados, estudiantes, amas de casa, menores de 6 años y personas con discapacidades. Esta categorización permite entender mejor los factores que influyen en la inactividad laboral, lo cual es crucial al estudiar la relación entre inactividad y desocupación dentro de los hogares.

En términos de las frecuencias y proporciones, una parte significativa de la muestra no tiene categoría asignada, con un 45.60% de los registros correspondientes al valor 0.0. Este alto porcentaje se corresponde con las personas que se encuentran en actividad, tienen asignado el valor 0 porque no les corresponde una clasificación de inactividad.

El grupo de estudiantes representa el mayor porcentaje dentro de las categorías válidas de inactividad, con un 22.14%. Este dato es esperado, dado que los estudiantes suelen estar fuera del mercado laboral debido a sus estudios. La categoría de amas de casa ocupa el tercer lugar con un 10.72%, reflejando a aquellos que no participan en el mercado laboral debido a sus responsabilidades

domésticas. Su presencia en un hogar podría influir en las dinámicas económicas y en la proporción de personas activas que pueden estar buscando empleo.

El 10.11% de la población corresponde a jubilados y pensionados, lo que indica que una porción significativa de la muestra está compuesta por adultos mayores que han dejado el mercado laboral. En cuanto al grupo de menores de 6 años, que constituye el 8.13% de la muestra, aunque no participan en el mercado laboral, es relevante debido a la dependencia económica que generan en los hogares, lo cual puede influir en las decisiones económicas de las familias.



Por otro lado, las categorías de discapacitados (0.73%) y rentistas (0.23%) tienen proporciones pequeñas, pero no por ello menos importantes. Las personas con discapacidad pueden enfrentar barreras adicionales para participar en el mercado laboral, mientras que los rentistas podrían tener un ingreso pasivo que los mantiene fuera del ámbito laboral activo.

En conclusión, la variable categoría de inactividad ofrece una visión detallada de los grupos inactivos dentro de los hogares y proporciona contexto adicional para comprender cómo estos grupos impactan en la economía familiar y en las tasas de desocupación.

Parte II: Clasificación y regularización

2) Elegimos el parámetro de regularización a través de validación cruzada dividiendo el conjunto de entrenamiento en k-folds para optimizar el hiper parámetro. Se entrena el modelo en k -1 de los subconjuntos y validamos en el restante. Utilizaremos k = 10. Esto se repite rotando los conjuntos y calculamos el promedio del ECM como métrica de desempeño para elegir el λ que lo minimiza. No se utiliza el conjunto de prueba hasta la evaluación de rendimiento final para evitar caer en *overfitting*.

3) Las implicancias de usar un *k* pequeño, por ejemplo < 5, son que cada *fold* va a tener muchos datos pero menos variabilidad entre la partición de entrenamiento y validación, cada fold va a ser más general y no tan particular. No vas a tener tantas combinaciones de datos. Con un *k* muy grande, como sería el caso de *leave one out* (*k* = n), la varianza aumenta. En consecuencia, los *outputs* están altamente correlacionados de manera positiva entre sí. En el caso de *leave one out*, se crean n *folds* (cada observación es un subconjunto) y se entrena en n-1 *folds* y se testea en la observación restante.

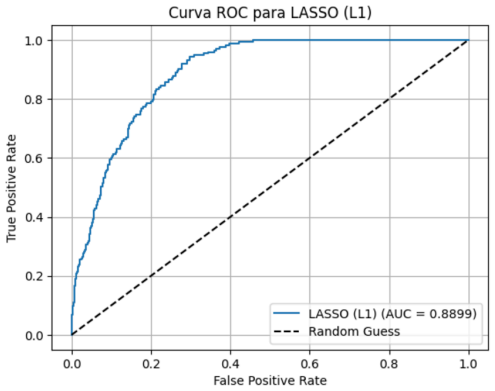
4) Se realizó una regresión logística con penalidad $\lambda = 1$ con LASSO (L1) y Ridge (L2) para ambos años, es decir, 2004 y 2024.

Cuando se realizó la regresión logística con LASSO en la base de datos de 2004, se observó que la matriz de confusión era bastante mala, ya que predijo que 168 de los 173 desocupados estaban ocupados (falso negativo). Sin embargo, como la mayoría de los datos estaban ocupados, el modelo

tiene un *accuracy* de 0.9264 y un AUC de 0.8899. Por esta misma razón, la curva de ROC dio muy buena. Por lo tanto, este es un muy mal modelo para predecir la desocupación.

LASSO - 2004:

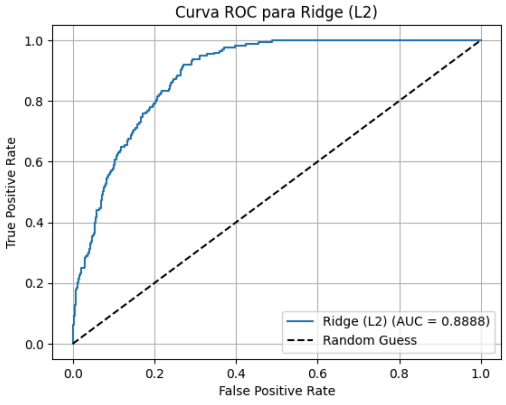
```
Resultados para penalización: LASSO (L1)
Matriz de Confusión:
[[2121    1]
 [ 168    5]]
Accuracy: 0.9264
AUC: 0.8899
              precision    recall  f1-score   support

           0       0.93      1.00      0.96      2122
           1       0.83      0.03      0.06       173

    accuracy                           0.93      2295
   macro avg       0.88      0.51      0.51      2295
weighted avg       0.92      0.93      0.89      2295
```



Al igual que con LASSO, la matriz de confusión de Ridge también obtuvo un alto porcentaje de falsos negativos (169 de 173). En consecuencia, su *accuracy* (0.9259) y su AUC (0.8888), al igual que la curva de ROC, dieron muy buenos resultados. De todas formas, es un mal modelo para predecir desocupados, ya que predice a casi todos los datos como ocupados.

Ridge - 2004:

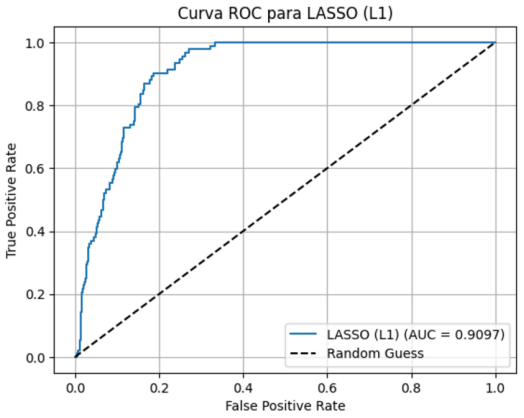
```
Resultados para penalización: Ridge (L2)
Matriz de Confusión:
[[2121    1]
 [ 169    4]]
Accuracy: 0.9259
AUC: 0.8888
              precision    recall  f1-score   support

           0       0.93      1.00      0.96      2122
           1       0.80      0.02      0.04       173

    accuracy                           0.93      2295
   macro avg       0.86      0.51      0.50      2295
weighted avg       0.92      0.93      0.89      2295
```



Algo similar ocurrió cuando se realizó la regresión logística en la base de datos de 2024. Por un lado, en la regresión con LASSO, se puede observar que en la matriz de confusión hay una muy alta tasa de error, donde de los 92 desocupados a todos los clasifica como ocupados. Teniendo en cuenta que el modelo predice a la mayoría como ocupados, tiene sentido que se reporte un *accuracy* de 0.9461 y un AUC de 0.9097, ya que la mayoría de los datos son ocupados. Además, la curva de ROC, si bien no está sobre los ejes, es buena. Pero, al considerar que la mayoría de los datos son ocupados y predice que casi todos son ocupados, tiene sentido que sea buena. Sin embargo, es un muy mal modelo de predicción para desocupados. Por el otro, en la regresión con Ridge sucedió algo similar a lo que sucedió con LASSO. Los valores de la matriz de confusión fueron iguales a los de LASSO, es decir, predijo que casi todos los datos eran ocupados. Por lo tanto, tuvo un *accuracy* de 0.9466 y un AUC de 0.9066, ya que la mayoría de los datos eran ocupados. Asimismo, la curva de ROC también dio muy buena, por la misma razón que la de LASSO fue buena. En conclusión, ambos modelos son muy malos para predecir si una persona está desocupada.

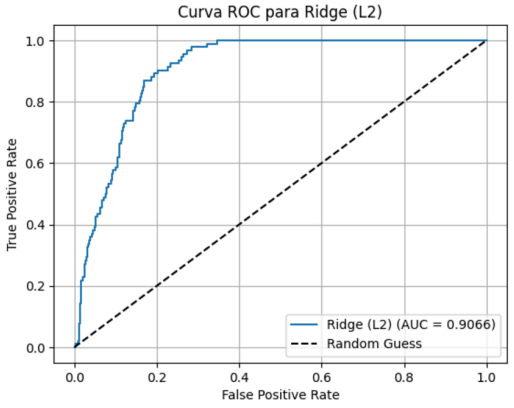
LASSO - 2024:

```
Evaluación para 2024 - LASSO (L1)
Resultados para penalización: LASSO (L1)
Matriz de Confusión:
[[1719    6]
 [  92    0]]
Accuracy: 0.9461
AUC: 0.9097
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1725
           1       0.00      0.00      0.00        92

    accuracy                           0.95      1817
   macro avg       0.47      0.50      0.49      1817
weighted avg       0.90      0.95      0.92      1817
```



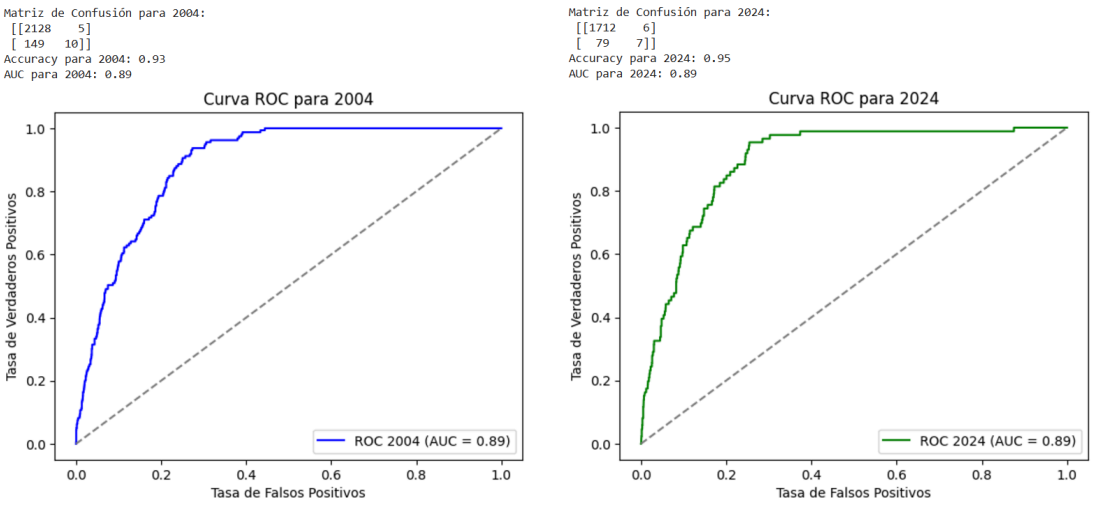
Ridge - 2024:

```
Evaluación para 2024 - Ridge (L2)
Resultados para penalización: Ridge (L2)
Matriz de Confusión:
[[1720    5]
 [  92    0]]
Accuracy: 0.9466
AUC: 0.9066
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1725
           1       0.00      0.00      0.00        92

    accuracy                           0.95      1817
   macro avg       0.47      0.50      0.49      1817
weighted avg       0.90      0.95      0.92      1817
```



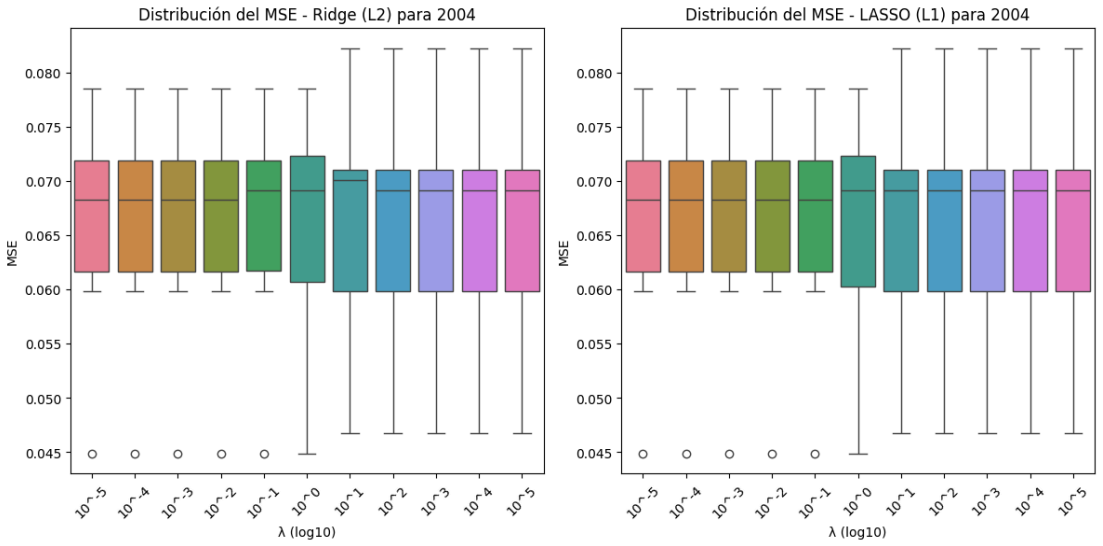
Para comparar de manera correcta los modelos en función de si tienen o no regularización, corregimos y volvimos a correr el modelo de regresión logística del TP3. Obtuvimos resultados diferentes a la primera vez ya que implementamos las *dummies* y quitamos la variable "Estado" de la matriz de regresoras para evitar la colinealidad perfecta del TP3.

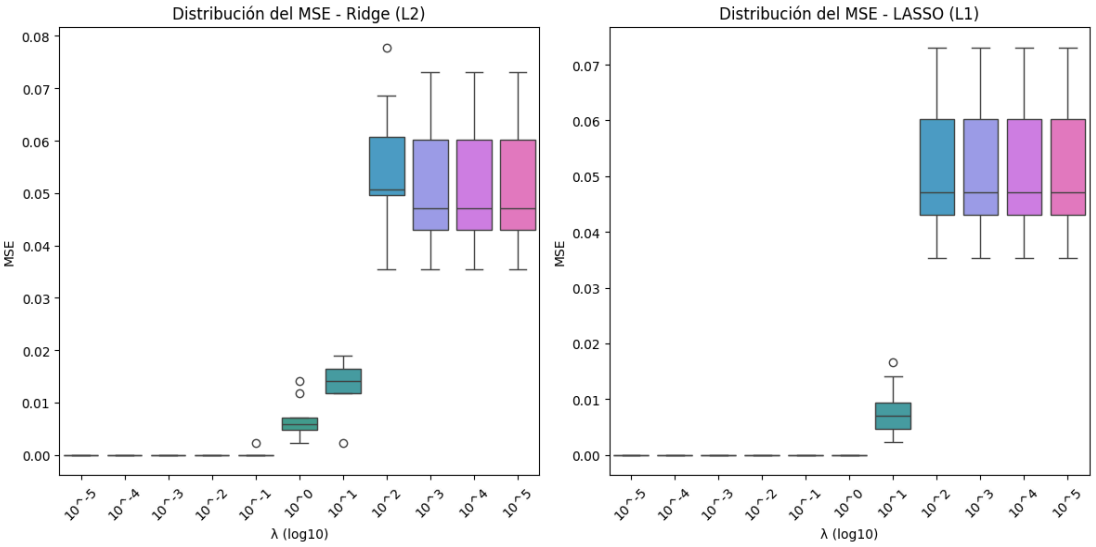
Para ambos años (2004 y 2024), tanto los modelos de regresión logística del TP3 como los modelos con regularización son malos prediciendo. En los primeros es posible observar que su matriz de confusión tiene una alta tasa de error (tan solo 10 y 7, respectivamente, observaciones correctamente clasificada), aunque tienen un alto *accuracy* (0.93 y 0.95), un alto AUC (0.89 y 0.89) y buenas curvas de ROC. Comparando estos resultados con los modelos regularizados analizados anteriormente, es posible observar que no hubo una mejora en el modelo. Es más, se podría decir que predijeron aún peor.



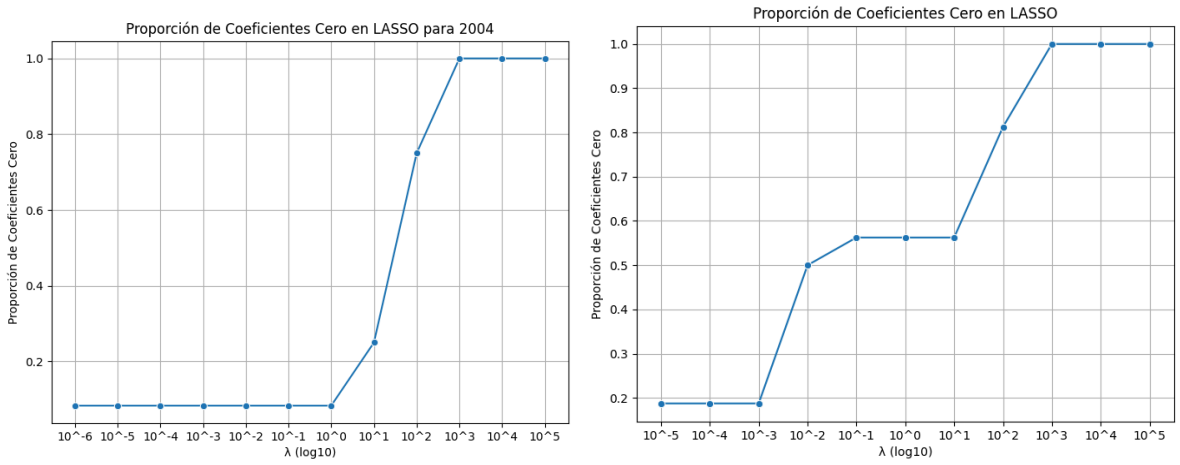
5) Se realizó un barrido en $\lambda = 10^{n} con\, n \in [-5, 5]$ y, para seleccionar el λ óptimo, se realizó un 10-*fold cross-validation* para ambos modelos (LASSO y Ridge) de ambos años y se seleccionó aquel λ con el menor error cuadrático medio (MSE). Para el 2004, se seleccionó $\lambda = 1e^{-5}$ para el modelo de Ridge y $\lambda = 1$ para LASSO. Para el 2024, se seleccionó $\lambda = 1e^{-5}$ para ambos modelos.

Además, se realizaron *boxplots* para poder representar gráficamente la distribución del MSE de cada modelo para cada año. En ellos es posible observar que a menor el valor de λ, más parece tener una distribución normal el MSE, para ambos modelos.





Luego, se realizó un gráfico que mostrara la proporción de coeficientes cero en LASSO para ambos años. En el 2004, es posible observar que una vez que $\lambda \geq 10^1$, la proporción de coeficientes aumenta hasta llegar a 1 a partir de $\lambda \geq 10^3$.



6) Al correr el modelo de regresión logística con LASSO penalizando con el λ óptimo (1 para el 2004 y 1e-05 para 2024), las variables descartadas en ambos casos es el intercept. Esto podría sugiere que la inclusión de todas las variables es necesaria para predecir desocupación (porque minimizan el error cuadrático medio), pero también nos hace pensar o que no incluimos suficientes variables en el modelo para que pueda penalizar y seleccionar sin perder poder predictivo, o que λ no es suficientemente grande para forzar una mayor regularización. Si aumentamos un poco el valor de λ hasta 1e1 para ver cuales son las primeras variables descartadas podemos observar que:

- Para el 2004, se le asigna coeficiente cero al intercepto, la variable dummy "MENOR" y la variable creada por nosotras "DENSIDAD"
- Para el 2024, se le asigna coeficiente cero al intercepto, y a "DENSIDAD"

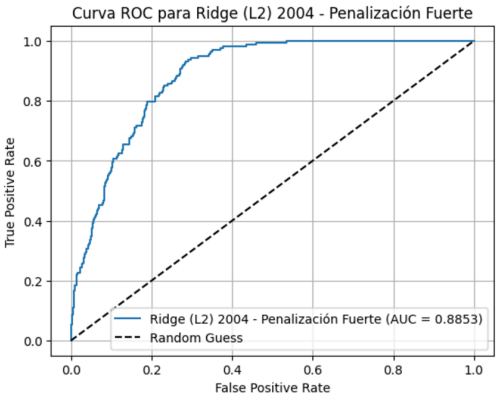
Esperábamos que se descarte el intercepto ya que normalizamos las variables para aplicar los modelos, sumado a que es difícil de interpretar en una regresión logística. Por otro lado, si bien la densidad familiar/habitacional es un indicador de nivel socioeconómico, no es un buen predictor necesariamente de desocupación.

7) Realizamos un modelo de regresión logística con regularización LASSO y otro con Ridge, ambos con una penalización de $\lambda = 1e$.

Luego de haber realizado el análisis, es posible concluir que ningún modelo es bueno para predecir la desocupación, sin importar el método de regularización ni el año. Esto se debe a que la matriz de confusión de los cuatro modelos predice todos los desocupados como ocupados, es decir, está prediciendo que todas las personas están ocupadas. Esto explica por qué el *accuracy* y el AUC dan valores tan altos (mayores a 0.85). Además, las curvas de ROC en todos los casos son buenas. Sin embargo, no clasifica correctamente ninguna observación de desocupado. Por lo tanto, ninguno de los modelos es bueno.

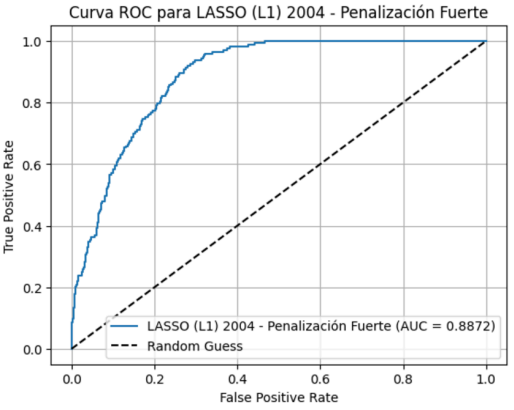
Ridge - 2004

```
Resultados para penalización: Ridge (L2) 2004 - Penalización Fuerte
Matriz de Confusión:
[[2122    0]
 [ 173    0]]
Accuracy: 0.9246
AUC: 0.8853
              precision    recall  f1-score   support

           0       0.92      1.00      0.96      2122
           1       0.00      0.00      0.00       173

    accuracy                           0.92      2295
   macro avg       0.46      0.50      0.48      2295
weighted avg       0.85      0.92      0.89      2295
```



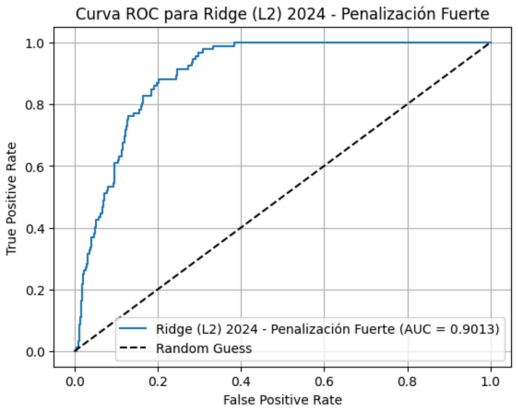
LASSO - 2004

```
Resultados para penalización: LASSO (L1) 2004 - Penalización Fuerte
Matriz de Confusión:
[[2122    0]
 [ 173    0]]
Accuracy: 0.9246
AUC: 0.8872
              precision    recall  f1-score   support

           0       0.92      1.00      0.96      2122
           1       0.00      0.00      0.00       173

    accuracy                           0.92      2295
   macro avg       0.46      0.50      0.48      2295
weighted avg       0.85      0.92      0.89      2295
```



Ridge - 2024

```
Resultados para penalización: Ridge (L2) 2024 - Penalización Fuerte
Matriz de Confusión:
[[1722    3]
 [  92    0]]
Accuracy: 0.9477
AUC: 0.9013
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1725
           1       0.00      0.00      0.00        92

    accuracy                           0.95      1817
   macro avg       0.47      0.50      0.49      1817
weighted avg       0.90      0.95      0.92      1817
```



LASSO - 2024

```
Resultados para penalización: LASSO (L1) 2024 - Penalización Fuerte
Matriz de Confusión:
[[1722    3]
 [  92    0]]
Accuracy: 0.9477
AUC: 0.9087
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1725
           1       0.00      0.00      0.00        92

    accuracy                           0.95      1817
   macro avg       0.47      0.50      0.49      1817
weighted avg       0.90      0.95      0.92      1817
```