

**Trabajo práctico número 4**

**Carrera**

Licenciatura en Ciencias del comportamiento

**Asignatura**

Ciencia de datos

**Autoras**

Luciana Crupnik

Josefina Laborda

Justina Reinke

**Profesor**

[Ignacio Spiousas](mailto:ispiousas@udesa.edu.ar)

Maria Noelia Romero

Realizado un análisis previo de la Encuesta Permanente de Hogares (EPH) y la desocupación, se llevará a cabo un nuevo estudio sobre el desempleo en relación con cuestiones del hogar.

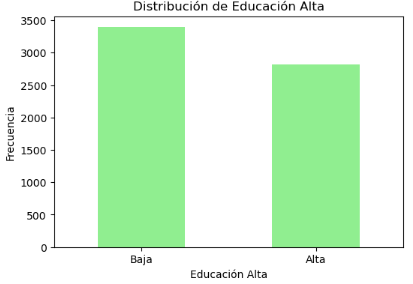
**Parte 1: Análisis de la base de hogares y tipo de ocupación**

En base al análisis de la base de datos “Hogar”, se seleccionaron algunas variables relevantes para predecir la desocupación. Entre ellas, se encuentra el Estado de Actividad, la cual clasifica a las personas como ocupadas, desocupadas, inactivas o menores de 10 años. A su vez, el Nivel Educativo, la cual refleja la relación entre educación y empleo, en donde niveles más bajos pueden asociarse con mayores tasas de desocupación. Por otro lado, las variables Región y Aglomerado muestran diferencias regionales que pueden influir en las oportunidades de empleo. El Ingrato total familiar y per cápita también pueden ayudar a predecir el empleo, en donde los hogares con menores ingresos podrían estar más asociadas con situaciones de desocupación. La cobertura médica, las condiciones habitacionales, la condicion laboral previa, la busqeuda activa de empleo, la edad, el sexo y el tipo de hogar, también pueden considerarse varibales relevantes a la hora de realizar una predicción.

Para comenzar con el análisis, se seleccionaron únicamente las observaciones que corresponden a los aglomerados de Ciudad de Buenos Aires o de Gran Buenos Aires, y posteriormente se unificaron las bases de la encuesta individual de cada año y con las bases de la encuesta de hogar.

En una primera instancia, se llevó a cabo una limpieza de la base de datos. El criterio tomado en cuenta fueron los “Missing Values”. Entre algunas de las variables con datos faltantes, se encuentran “ch14”, “ch15\_cod”, “ch16\_cod”, “pp03c” y “pp03d”. Por lo tanto, se decidió reemplazar estos valores por la mediana o por la moda, según si son datos numéricos o datos categóricos, respectivamente. Sin embargo, luego de realizar este procedimiento, se identificaron algunas variables que todos sus observaciones eran datos faltantes, tales como "pp09c\_esp", "idecocur", "pdecocur", "idecindr", "pdecindr", "idecifr\_x", "pdecifr\_x", "ideccfr\_x", "pdeccfr\_x", "pondih". De todas formas, se verificó que esto no esté relacionado a un problema en el código, sino que desde un principio no contenían valores. Por lo tanto, se tomó la decisión de reemplazarlas a 0. En el caso de “pondih”, al unificar los dataframes, se crearon dos nuevas variables como “pondih\_x” y “pondih\_y”, los cuales contenían valores numéricos, así que no se perdieron datos relevantes al convertir sus observaciones a 0. En una segunda instancia, se verificaró las variables que contenían valores negativos en la base de datos individual del trabajo práctico anterior, y en caso de que suceda, se eliminaron esas filas.

Una vez finalizada la limpieza de datos, se propusieron tres nuevas variables que sean relevantes para predecir a los individuos desocupados. La primera variable definida se nombró como “Educación alta”, la cual intenta identificar aquellos que han alcanzado un nivel de educación alta, tomando los siguientes valores: 1 si el individuo ha completado al menos el nivel secundario (o más) o 0 si no ha completado el nivel secundario o no tiene educación formal. Los resultados muestran que aproximadamente el 45% de los individuos en la muestra tienen un nivel educativo alto (al menos secundario completo), pero que el 55% tiene niveles educativos bajos o interrumpidos antes de completar la secundaria. La relación entre la educación y la desocupación es fuerte, ya que las personas con niveles educativos más bajos suelen tener menos competencias y habilidades específicas requeridas en el mercado laboral. Además, sin al menos un nivel secundario completo, las opciones laborales se restringen. Por lo tanto, incorporar esta variable puede ayudar a identificar tendencias de desocupación.

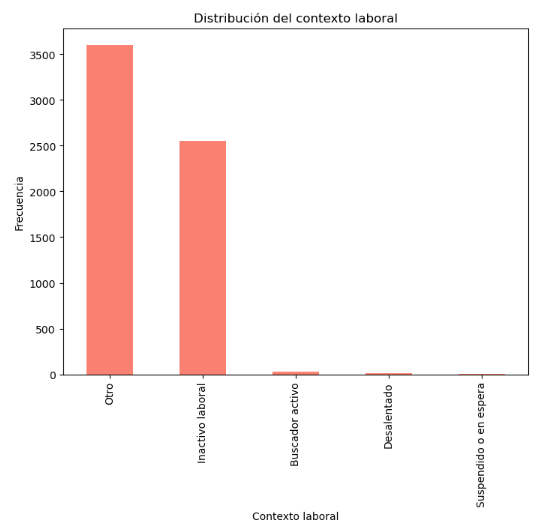


*Gráfico 1: Distribución de la variable “Educación Alta”*

En segundo lugar, se definió la variable “Contexto laboral”, y su función es clasificar a las personas, teniendo en cuenta las categorías: Motivos para no buscar empleo (pp02e), si han buscado trabajo en los últimos 12 meses (pp02h) y si han trabajado en los últimos 12 meses (pp02i). La clasificación divide a los datos de la siguiente manera:

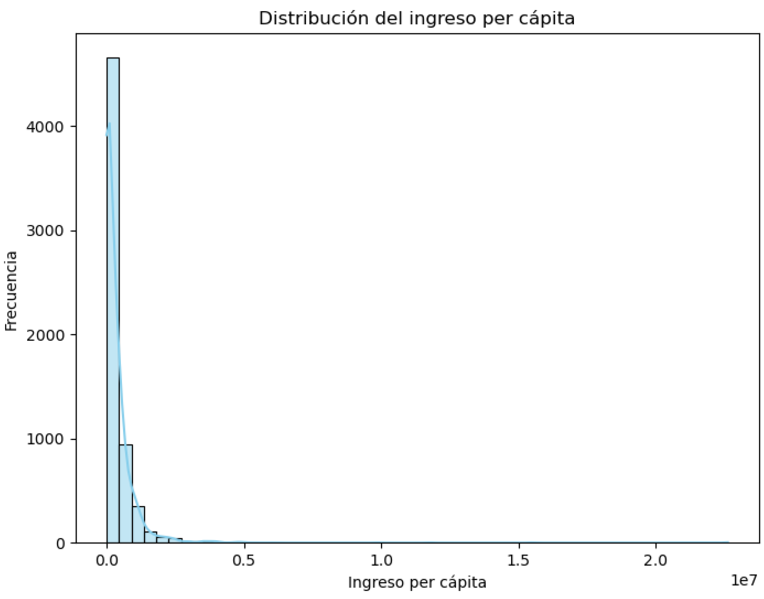
1. Buscador activo: Ha buscado empleo en los últimos 12 meses y no tiene trabajo asegurado (pp02e ≠ 2).
2. Desalentado: No buscó empleo porque se cansó o considera que hay poco trabajo (pp02e = 3 o 4).
3. Suspendido o en espera: No buscó empleo porque está suspendido o tiene trabajo asegurado (pp02e = 1 o 2).
4. Inactivo laboral: No trabajó ni buscó empleo en los últimos 12 meses (pp02h = 2 y pp02i = 2).

Entre los resultados, se obtuvo que el 58,04% incluye personas que no encajan con las definiciones especificadas de la categoría, el 41.12% de las personas no buscaron trabajo en los últimos 12 meses, el 0,53% buscaron trabajo activamente en los últimos 12 meses, último, el 0,19% no buscaron empleo porque se cansaron o porque perciben que no hay trabajo disponible, y por último, el 0,11% incluye personas que no buscaron trabajo porque están suspendidas temporalmente o tienen trabajo asegurado.



*Gráfico 2: Distribución de la variable “Contexto Laboral”*

En tercer lugar, la variable “Ingreso Per Cápita” se definió a partir de la división entre “Ingreso Total Familiar” (itf) y el número de miembros del hogar (ix\_tot). Los resultados reflejan el ingreso promedio por miembro del hogar, lo que permite comparar el nivel de ingresos ajustado por el tamaño del hogar. En los resultados, se observó que la media del Ingreso Per Cápita fue de $361,988.00, con un máximo de Ingreso de $22,624,670.00. Por otro lado, la desviación estándar de esta variable es de $810,463.70, mostrando una gran dispersión entre los datos. Esto refuerza la idea de desigualdad entre los hogares. Por lo tanto, el Ingreso Per Cápita es una métrica que se relaciona fuertemente con la desocupación. Los hogares con bajo ingreso per cápita pueden reflejar la falta de empleo en miembros del hogar en edad laboral o subempleos en donde los ingresos no alcanzan niveles suficientes para sostener al hogar.



*Gráfico 3: Distribución de la variable “Ingreso por Cápita”*

**Parte 2: Clasificación y regularización**

El objetivo de esta sección es predecir si una persona se encuentra desocupada, utilizando distintas variables de la encuesta EPH que contemplan tanto características individuales como aspectos relacionados con el hogar del encuestado.

Para ello, lo primero que hicimos fue dividir la base de datos inicial en dos subconjuntos correspondientes a los años 2004 y 2024, basándonos en la variable ano4. Posteriormente, filtramos la base para quedarnos únicamente con las observaciones en las que los encuestados respondieron su condición de actividad. Además, agregamos una nueva columna llamada desocupado, que identifica si la persona se encontraba desocupada (estado igual a 2). A continuación, dividimos las bases respondieron\_2004 y respondieron\_2024 en conjuntos de entrenamiento y prueba, asignando el 70% de los datos al conjunto de entrenamiento y el 30% al conjunto de prueba. Para realizar esta partición, utilizamos el comando train\_test\_split y establecimos la semilla 101 para garantizar reproducibilidad en los resultados.

En cada conjunto, definimos la variable desocupado como la variable dependiente y utilizamos el resto de las variables como variables independientes. Por último, añadimos una columna de unos a las matrices X. El resultado final de este proceso fue el siguiente: para el año 2004, el conjunto de entrenamiento quedó conformado por 303 observaciones y el conjunto de prueba por 130 observaciones, mientras que para el año 2024, el conjunto de entrenamiento incluyó 3966 observaciones y el conjunto de prueba 1701 observaciones.

Para seleccionar el mejor valor de lambda en los modelos de regresión Ridge para los años 2004 y 2024, utilizamos validación cruzada con el método GridSearchCV. Este enfoque nos permitió explorar un rango de valores de lambda ( = 0.01, 0.1, 1, 10, 100) y evaluar el rendimiento de cada modelo en términos del error cuadrático medio negativo. La validación cruzada se realizó con 5 particiones, lo que implica que en cada iteración se entrena el modelo con 4 particiones y se valida en la restante. El procedimiento comenzó estandarizando las variables predictoras en ambos años lo cual es esencial para garantizar que la regularización actúe de manera uniforme sobre todas las variables. Luego, entrenamos el modelo Ridge con cada valor de lambda en los datos de entrenamiento, evaluando su rendimiento en las particiones de validación. El valor de lambda que minimizó el error promedio en las particiones se seleccionó como el mejor.

Para 2004, el mejor lambda fue igual a 100, y al entrenar el modelo final con este valor, obtuvimos un error cuadrático medio (MSE) en el conjunto de prueba de 0.0667. Para 2024, el mejor lambda fue igual a 100, y el MSE en el conjunto de prueba fue 0.04788.

Es importante destacar que no utilizamos el conjunto de prueba para seleccionar lambda. Esto se debe a que el conjunto de prueba debe reservarse exclusivamente para evaluar el desempeño final del modelo y así estimar su error cuadrático medio estimado fuera de muestra. Usar el conjunto de prueba para la selección de lambda podría introducir overfitting, ya que el modelo se ajustaría excesivamente a las particularidades de ese conjunto, perdiendo capacidad de generalización.

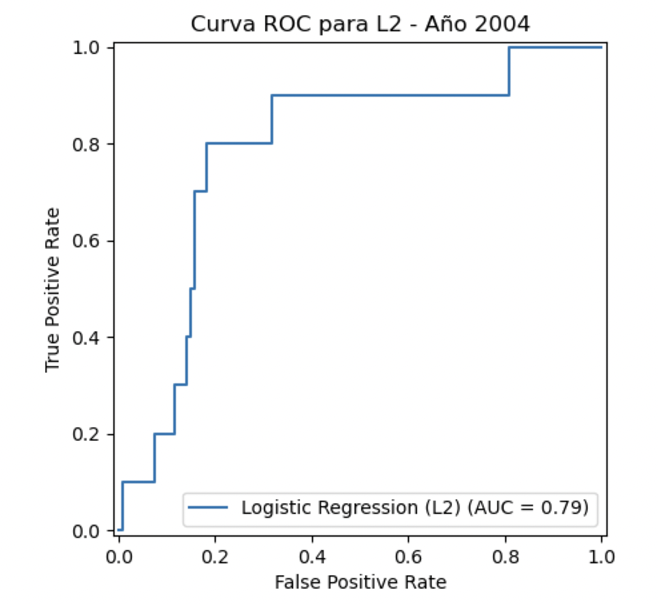
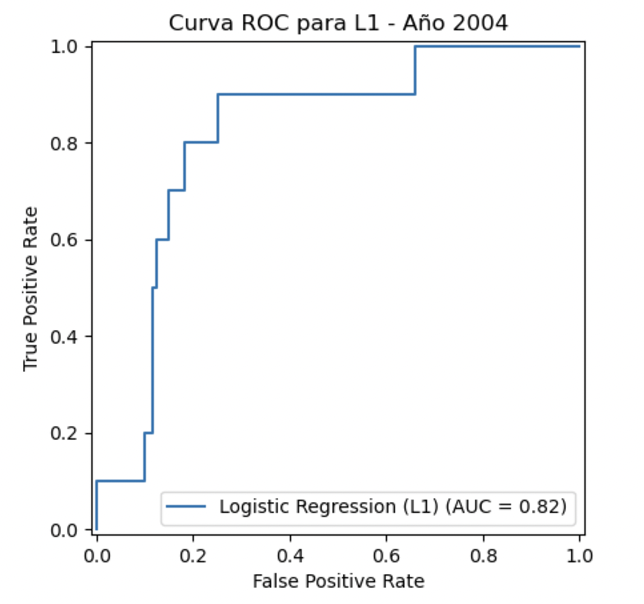
Luego, se utilizó el método de validación cruzada Leave-One-Out (LOOCV), donde k=n (el número total de observaciones), para estimar el error cuadrático medio (MSE) y seleccionar el nivel óptimo de regularización (). LOOCV es una técnica basada en la ley de los grandes números, que asegura que, al aumentar las particiones, el promedio de los errores de predicción se aproxima al valor verdadero del error de pronóstico fuera de la muestra. En este método todas las observaciones son usadas tanto para entrenamiento como para validación.El procedimiento se implementó ajustando un modelo Ridge con diferentes valores de sobre el conjunto de entrenamiento de 2004. En cada iteración, se entrenó el modelo con n−1 observaciones y se validó utilizando la muestra restante, calculando el MSE para cada valor de . Posteriormente, se promediaron estos MSE para determinar el valor óptimo de , minimizando el error de predicción. Con LOOCV, se obtuvo un óptimo de 100, asociado a un ECM promedio de 0.0477 en el conjunto de validación. El modelo final, ajustado con este , fue evaluado en el conjunto de prueba, obteniendo un ECM de 0.0666.

La elección de k=n en LOOCV permite maximizar el uso de los datos para entrenar y evaluar el modelo. Sin embargo, tiene implicancias importantes. Por un lado, al utilizar cada observación como muestra de prueba una vez, y de entrenamiento n−1 veces, se obtienen estimaciones más precisas del error fuera de la muestra, reduciendo el sesgo en la selección de complejidad del modelo. Por otro lado, el costo computacional es considerablemente alto, especialmente en conjuntos de datos grandes, ya que el modelo debe ajustarse *n* veces. El número de particiones (k) en la validación cruzada influye directamente en el balance entre sesgo, varianza y costo computacional. Un k pequeño, como 2, ofrece particiones grandes para el entrenamiento, pero podría generar estimaciones menos confiables del error fuera de la muestra. Por el contrario, un k grande, como en LOOCV, minimiza el sesgo, pero incrementa la varianza, además de requerir un esfuerzo computacional mayor.

El uso de validación cruzada en este contexto no solo ayuda a seleccionar el nivel óptimo de complejidad del modelo (), sino también a evitar el problema de sobreajuste (overfitting). Modelos demasiado complejos tienden a ajustar perfectamente los datos de entrenamiento, pero fallan en predecir a nuevos datos de testeo. La curva de error de validación cruzada permite identificar el punto en el que el MSE de prueba es mínimo, proporcionando una guía para equilibrar la complejidad y el performance del modelo.

Se implementó un modelo de regresión logística utilizando las penalizaciones L1 (LASSO) y L2 (Ridge) con un parámetro de regularización λ = 1, empleando la opción penalty para ajustar los modelos. En el caso de la penalización L1 para el año 2004, los resultados obtenidos muestran un desempeño adecuado del modelo, aunque con ciertas limitaciones en la predicción de verdaderos positivos. La matriz de confusión identificó correctamente 119 casos negativos y 1 positivo. Esto resultó en 1 falso positivo y 9 falsos negativos. Esto sugiere que el modelo tiene una alta capacidad para identificar casos negativos, pero una baja capacidad para distinguir casos positivos. El valor de precisión general alcanzó un 92.31 %, lo que refleja un desempeño aceptable en la clasificación de la clase mayoritaria (negativos). Sin embargo, el valor del AUC (Área Bajo la Curva) de 0.8183 indica que el modelo tiene una capacidad moderada para discriminar entre clases

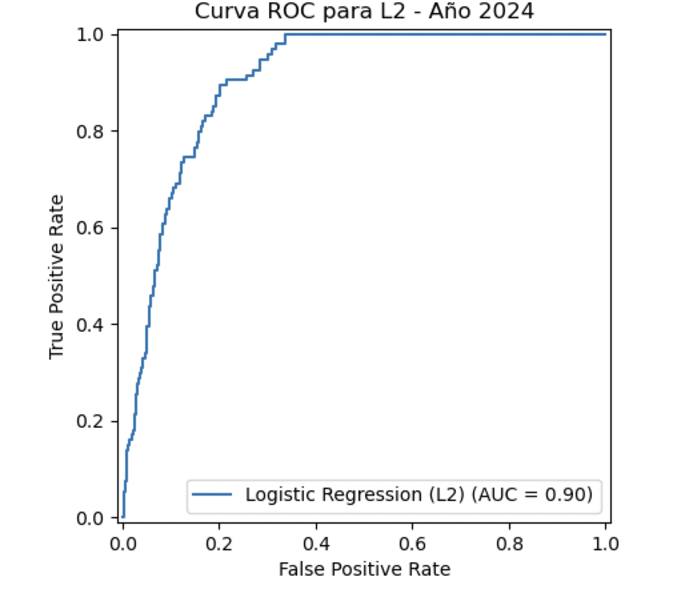
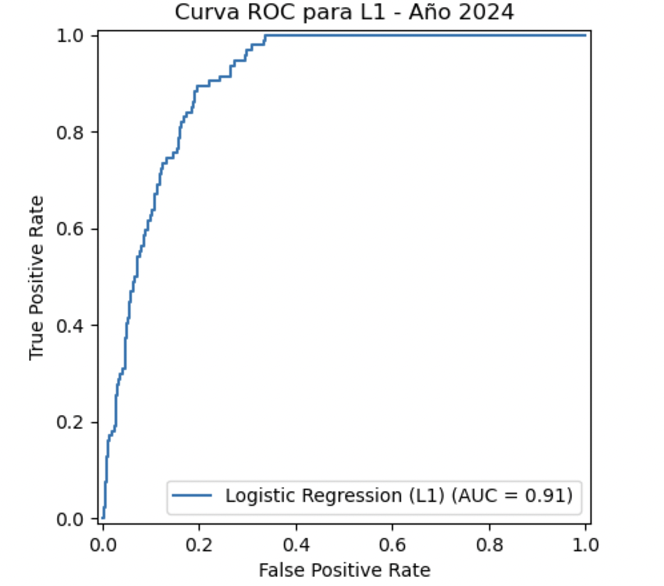
Para la penalización L2 aplicada al mismo año, se observaron resultados similares a los de la penalización L1. La matriz de confusión identificó correctamente 119 casos negativos y 1 positivo, con 1 falso positivo y 9 falsos negativos. El valor de precisión global fue del 92.31 %, igual que en el caso de L1. Sin embargo, el AUC obtenido fue de 0.7883, lo que indica una capacidad ligeramente menor para discriminar entre clases en comparación con la penalización L1.



*Figuras 5 y 6: Curva ROC para penalización L1 y L2 del año 2004.*

Por otro lado, para la penalización L1 aplicada al año 2024, los resultados muestran un buen desempeño del modelo. La matriz de confusión revela que el modelo clasificó correctamente 1604 casos negativos y 0 positivos, con 3 falsos positivos y 94 falsos negativos. Esto indica una alta precisión en la identificación de la clase negativa, pero una incapacidad para clasificar correctamente los casos positivos. El valor del AUC fue de 0.9050, lo que refleja una buena capacidad para discriminar entre clases. La precisión global del modelo fue del 94.3 %, indicando un buen rendimiento general. Sin embargo, la incapacidad para identificar correctamente casos positivos (clase minoritaria) sugiere que el modelo podría estar influido por el desbalance de clases.

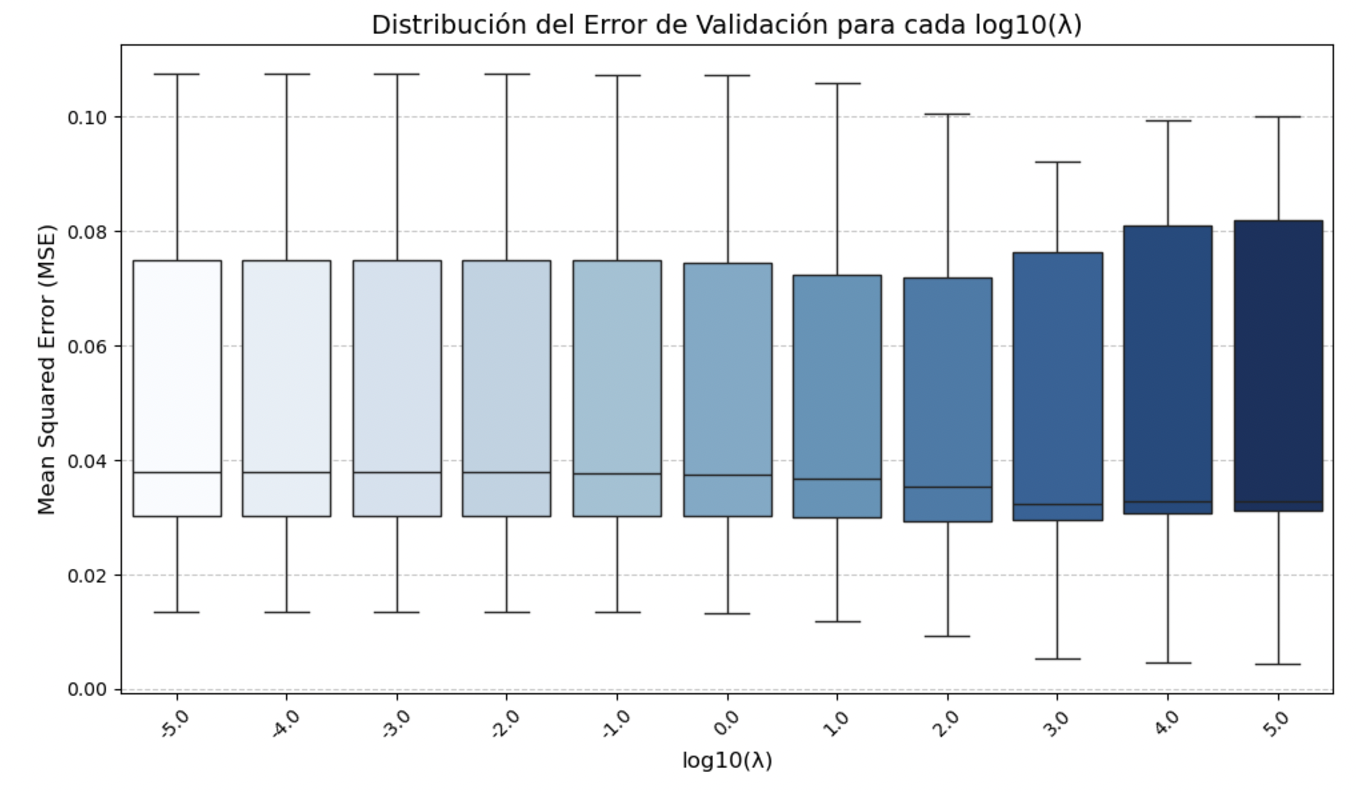
De la misma forma, la penalización L2 fue aplicada al año 2024, mostrando un rendimiento muy parecido al modelo con penalización L1. En este caso, la matriz de confusión identificó correctamente 1605 casos negativos y 0 positivos, con 2 falsos positivos y 94 falsos negativos. El AUC fue de 0.9048, lo que indica una buena capacidad para discriminar entre clases, mientras que la precisión global del modelo fue del 94.36 %, reflejando un buen desempeño general. Sin embargo, al igual que con la penalización L1, la ausencia de verdaderos positivos sugiere que el modelo podría estar influido por el desbalance de clases.

**

*Figura 6 y 7: Curva ROC para penalización L1 y L2 de 2024.*

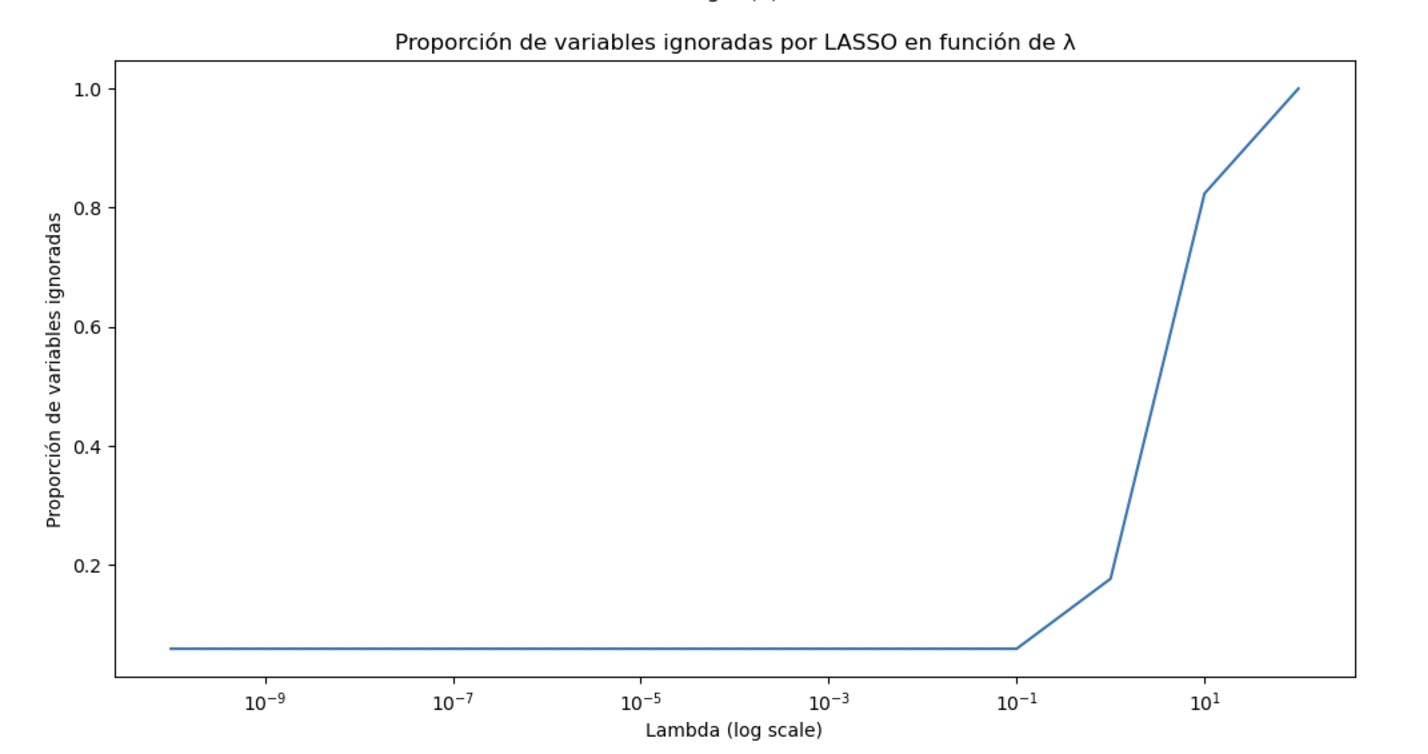
Al comparar los resultados obtenidos con regularización respecto al modelo de regresión logística sin regularización realizado en el trabajo práctico 3, no se observa una mejora en la performance del modelo. En el modelo de regresión logística sin regularización alcanzó una precisión del 99.5 % y un AUC de 0.9987, reflejando un buen desempeño con pequeños errores: un falso positivo y diez falsos negativos. En contraste, al implementar la regularización (penalizaciones L1 y L2) para ambos años, se evidenció que el modelo empeoró en varios aspectos clave. En cuanto a los errores de clasificación, no clasificó correctamente los casos positivos, resultando en una cantidad elevada de falsos negativos en todos los casos. El AUC mostró un valor menor a 0.91 para todos los casos, un valor bajo respecto al modelo de regresión logística sin regularización. En cuanto a la precisión, aunque con penalización el modelo indicó valores altos, ninguno superó al del trabajo anterior.

Luego, se realizó un barrido de valores de regularización en modelos de regresión logística con penalizaciones L1 (LASSO) y L2 (Ridge), para elegir el óptimo para ambos años. En la biblioteca scikit-learn la regularización se ajusta a través del parámetro , por lo que se estableció un rango de valores de expresado como donde varió entre -5 y 5. Para determinar el mejor desempeño, se utilizó validación cruzada con 10 particiones, evaluando cada valor de y seleccionando aquel que alcanzó el mayor ROC-AUC. Los valor óptimos de obtenidos fueron de 10.00 para Ridge 1.00 para LASSO. Para visualizar el error de predicción para cada , generamos un boxplot. Cada box corresponde a un valor de y contiene el error cuadrático medio (MSE) para cada partición. Se observa que los valores de log⁡10( entre -5 y - 2 presentan el menor error cuadrático medio (MSE) y su mediana es la más baja comparado a otros valores. A medida que log⁡10( aumenta, la variabilidad y el error cuadrático medio aumentan. Por lo tanto, los valores óptimos de son log⁡10 ] ya que parecen ser los óptimos para minimizar el MSE logrando el mejor equilibrio entre minimizar el error cuadrático medio (MSE) y un balance entre ajuste y regularización.



*Figura 8: Distribución del error de validación para cada log10(*

Luego, identificamos la proporción de variables ignoradas por el modelo en función de , es decir la proporción de variables para las cuales el coeficiente asociado es cero.Se generó un gráfico que muestra cómo LASSO elimina progresivamente variables a medida que incrementa. Valores altos de resultan en la exclusión de muchas variables, lo que puede perjudicar la capacidad predictiva del modelo al eliminar información relevante. En contraste, valores muy bajos de retienen demasiadas variables, lo que incrementa el riesgo de sobreajuste. Se determinó que un valor óptimo de se encuentra cerca de donde se logra un equilibrio adecuado: mantener variables relevantes y eliminar las menos significativas,

**

*Figura 9: Proporción de variables ignoradas por LASSO en función de*

Utilizando valor óptimo de para LASSO ( el modelo descarto las variables “ch04”, “ch06” y “ch07”, que corresponden a las caracteristicas de los miembros del hogar: sexo, años cumplidos, y vinculo actual (casado/separado/divorciado/viudo/soltero/unido). También descartó la variable 'ii7', que indica el régimen de tenencia, y variables características de la vivienda como 'iv3', que indica el material de los pisos, 'iv5' que indica si el techo tiene cielorraso / revestimiento interior, 'iv8' si la vivienda tiene baño o letrina y 'iv6', que indica la ubicación del acceso al agua. Descartó a su vez la variable 'ix\_men10', que indica la cantidad de menores de 10 años que habitan el hogar y la variable ‘vii2\_3' , que indica que otras personas que ayudan en las tareas de la casa. Por último, las que más sorprendieron fueron las variables 'ipcf\_x', que corresponde al monto del ingreso per cápita familiar y 'cat\_inac', que responde a la categoría de inactividad. A su vez nos sorprendió que haya descartado variables relacionadas a la educación, como 'nivel\_ed' y 'educacion\_alta', que indica el nivel de educativo alcanzado. Al comienzo del trabajo, creíamos que iban a ser fundamentales para realizar la predicción.

Por último, ajustamos los modelos de regresión logística con Ridge y Lasso con los óptimos encontrados para comparar ambos años. Para ello, se entrenaron ambos modelos sobre los conjuntos de datos de entrenamiento normalizados para cada año y se generaron predicciones sobre los datos de prueba. Para evaluar el desempeño, calculamos el MSE, y observamos qué modelo minimizó mejor el error entre las predicciones y los valores reales. Para el año 2004, obtuvimos un MSE de 0.0769 para Ridge y para Lasso, un valor ligeramente menor, de 0.0692. Esto puede deberse a la capacidad de Lasso para simplificar el modelo, eliminado variables irrelevantes o redundantes, mejorando así su capacidad predictiva para este conjunto de datos. Para el año 2024, obtuvimos también valores similares del MSE, en el caso de Ridge un valor de 0.0552 y de Lasso 0.0564. Estos resultados sugieren que Ridge fue más robusto en 2024, al mantener variables correlacionadas que pudieron ser descartadas por LASSO.

Finalmente, comparamos la selección de predictores que hizo LASSO en 2004 y 2024. Por un lado, en 2004 seleccionó 11 predictores, mientras que en 2024 seleccionó 15, manteniendo los mismos que en 2004, pero sumando 4 predictores más. Esto podría deberse a que en 2004 identificó relaciones claras y significativas entre estas variables y la desocupación, mientras que en 2024 considero factores adicionales relevantes para predecir la desocupación, lo que llevó al modelo a conservar más variables.