Informe TP4 - Ciencia de Datos

Daniela Chejfec, Esperanza Pereyra Iraola, Violeta Juliá

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

Ejercicio 1

La Encuesta Permanente de Hogares (EPH) permite obtener indicadores clave sobre el mercado laboral, como la tasa de desocupación. Para perfeccionar el análisis de esta tasa a nivel hogar, se identificaron variables relevantes de la base de datos. Entre ellas, se destaca *pondera* (peso de la observación), ya que permite interpretar las tasas de desempleo de manera representativa a nivel poblacional. Asimismo, las variables *region* y *aglomerado* son clave, dado que reflejan diferencias en las oportunidades laborales y el desarrollo económico según la ubicación geográfica. Por otro lado, el *IPCF* (ingreso per cápita familiar) resulta esencial para medir la vulnerabilidad económica de los hogares, mientras que el *tamaño del hogar* y la cantidad de integrantes ocupados o desocupados aportan información sobre la dependencia económica dentro del hogar. Además, variables relacionadas con la educación promedio del hogar, de estar disponibles, podrían proporcionar insights sobre la empleabilidad de los integrantes. Estas variables permitirán una construcción más robusta de las tasas de desocupación y un análisis más detallado de las dinámicas económicas de los hogares.

Ejercicio 2

Se descargaron las bases de EPH para los años 2004 y 2024 en los formatos .dta y .xlsx, respectivamente. En la base de 2004, se reemplazaron las etiquetas de la variable aglomerado por valores numéricos equivalentes, asignando 32 a la Ciudad de Buenos Aires y 33 a los Partidos del Gran Buenos Aires. Posteriormente, se filtraron ambas bases para incluir únicamente observaciones correspondientes a estos dos aglomerados.

Una vez filtradas, las bases de hogares e individuos fueron unidas para cada año utilizando las variables codusu y nro_hogar, aplicando un merge interno que asegura la correspondencia entre ambos niveles de análisis. Finalmente, se concatenaron las bases de 2004 y 2024, formando un único dataset denominado base_unificada, compuesto por 14,698 observaciones y 268 columnas. Este dataset consolidado fue exportado en formato .csv y contiene una variedad de variables relevantes, como identificadores únicos, localización geográfica y características socioeconómicas, que serán clave para los análisis posteriores.

Ejercicio 3

Se analizaron las columnas duplicadas al realizar el merge de las bases. Se encontraron columnas con sufijos _x y _y, que se compararon para decidir si conservar una sola. Si las columnas eran idénticas, se quedó con una sola; en caso contrario, se combinó tomando los valores no nulos. Tras esto, se eliminó del dataset las columnas redundantes. Luego, se aplicaron varios mapeos para convertir variables categóricas a numéricas, facilitando la manipulación de datos. A continuación, se limpiaron los valores de las columnas numéricas, asegurando que la edad, los ingresos y el sexo estuvieran dentro de rangos válidos. Finalmente, se filtró la base de datos para conservar únicamente las observaciones con valores válidos, reduciendo la cantidad de filas de 14,698 a 13,179. Como último paso, se eliminaron las columnas que contenían valores nulos en la base unificada para reducir la complejidad del análisis. Se guardó la base final limpia en un archivo CSV, que contiene 13,179 filas y 130 columnas, preparándose así para realizar los análisis estadísticos.

Ejercicio 4

Para analizar las condiciones socioeconómicas de los hogares, se calcularon tres variables. La primera variable es la dependencia económica, que se determina a partir de la proporción de personas fuera de la edad de trabajar en cada hogar. Para ello, se calculó la diferencia entre el total de integrantes del hogar y los mayores o iguales a 10 años, dividiendo el resultado por el total de integrantes. Esta medida resulta relevante porque refleja la carga económica sobre los miembros activos del hogar, lo cual puede estar vinculado a contextos de vulnerabilidad socioeconómica y a una menor integración al mercado laboral. La segunda variable es la proporción de personas sin secundaria completa, la cual se calcula al filtrar a las personas que no han completado la secundaria y luego dividir la cantidad de estas personas entre el total de integrantes del hogar. Esta proporción es crucial porque una alta proporción de personas sin secundaria completa está asociada con menores oportunidades laborales, lo que podría estar relacionado con un mayor riesgo de desocupación en el hogar. Por último, se calculó la proporción de subsidios sobre el ingreso total del hogar, lo que ayuda a entender la dependencia de los hogares respecto a subsidios sociales. Esta variable puede ser un indicio de vulnerabilidad económica y baja inserción laboral, lo cual también podría correlacionarse con mayores tasas de desocupación. Tras calcular estas variables, se verificaron los valores únicos de las columnas relevantes para asegurar que los cálculos fueran correctos, los cuales debían estar dentro del rango esperado de 0 a 1.

Ejercicio 5

Se calcularon estadísticas descriptivas para tres variables relevantes para la desocupación: el ingreso total del hogar, la edad y los subsidios sobre el ingreso. Para el ingreso total del hogar, el promedio es de $241548,18, con una desviación estándar de 823021,61, lo que sugiere una gran disparidad en los ingresos, con algunos hogares ganando cantidades muy altas. El valor mínimo es 0, lo que podría indicar hogares sin ingresos, y el máximo alcanza los 33937000, lo que refleja la presencia de casos extremos. Los percentiles 25 (450) y 75 (255000) indican que muchos hogares tienen ingresos bajos, pero con una pequeña proporción en tramos más altos.

Respecto a la edad, la media es de 35,02 años, con una desviación estándar de 22,82, lo que implica que hay una gran diversidad de edades entre las personas, desde muy jóvenes hasta mayores de 90 años. El valor mínimo es 1, lo cual parece un dato atípico, ya que es improbable que un niño de un año haya respondido la encuesta. El valor máximo es 96, lo que indica la presencia de personas mayores en la muestra. Los percentiles 25 y 75 muestran que la mayoría de las personas están en una franja de edad comprendida entre los 16 y los 52 años, lo que es consistente con la población en edad de trabajar.

Por último, para los subsidios sobre el ingreso, el promedio es de $1487,14, pero con una desviación estándar extremadamente alta de 13726,57, lo que sugiere que la mayoría de las personas no reciben subsidios significativos, aunque algunos casos extremos reportan montos elevados. La mediana y los percentiles 25 y 75 (todos en 0) indican que una gran parte de la población no recibe subsidios, mientras que el valor máximo de 300000 muestra casos aislados de subsidios muy altos. Este patrón refleja que los subsidios sociales son accesibles solo para un pequeño grupo, dejando a la mayoría sin ayudas.

Parte II: Clasificación y regularización

Ejercicio 1

Para predecir si una persona está desocupada, se realizó una división de la base de datos en conjuntos de entrenamiento y prueba. La división se llevó a cabo con el comando train_test_split de scikit-learn, asignando el 70% de los datos al conjunto de entrenamiento y el 30% al de prueba. Se utilizó la variable estado como la dependiente, codificada de forma binaria donde 1 indica que la persona está desocupada (estado=2) y 0 que no lo está.

El conjunto de entrenamiento quedó compuesto por 9225 observaciones y 134 variables predictoras (columnas independientes), mientras que el conjunto de prueba tiene

3954 observaciones con las mismas 134 variables. Esto asegura una proporción adecuada para construir un modelo robusto y evaluar su desempeño en datos que no fueron utilizados durante el entrenamiento. Además, esta partición respeta la semilla random_state=101, lo que garantiza la reproducibilidad del análisis. El procedimiento garantiza que el conjunto de entrenamiento tiene suficiente información para ajustar el modelo, mientras que el conjunto de prueba permite evaluar su capacidad de generalización en datos nuevos. De esta manera se evita el sobreajuste y permite medir el rendimiento del modelo de manera objetiva.

Ejercicio 2

La selección del valor óptimo de λ se realiza mediante validación cruzada, un método que divide el conjunto de entrenamiento en varios subconjuntos o folds. Durante este proceso, en cada iteración se reserva uno de los folds como conjunto de validación, mientras que los restantes se utilizan para entrenar el modelo. Esto permite evaluar diferentes valores de λ midiendo el error promedio en los conjuntos de validación a lo largo de todas las iteraciones. El valor de λ que minimice este error promedio será el seleccionado, ya que logra un balance óptimo entre el ajuste del modelo y su capacidad de generalización. No se utiliza el conjunto de prueba para seleccionar λ porque su propósito es evaluar la capacidad del modelo para generalizar en datos completamente nuevos, como se discutió en clases. Usarlo para la selección de λ comprometería su imparcialidad, ya que el modelo sería ajustado parcialmente con base en estos datos, aumentando el riesgo de sobreajuste y proporcionando un desempeño inflado en el conjunto de prueba. La validación cruzada, al usar únicamente el conjunto de entrenamiento, asegura que el conjunto de prueba permanezca reservado hasta la evaluación final, permitiendo medir objetivamente el desempeño del modelo en datos no vistos.

Ejercicio 3

En validación cruzada, el número de divisiones (k) afecta el balance entre la variabilidad de las estimaciones del error y el costo computacional del proceso. Si se elige un valor de k muy pequeño, cada fold contiene muchas observaciones para entrenamiento, lo que permite estimar el modelo con gran cantidad de datos. Sin embargo, el conjunto de validación será chico, lo que puede llevar a estimaciones del error con alta varianza y menor capacidad de generalización al no representar adecuadamente la diversidad del conjunto completo. Por otro lado, si k es muy grande, como en el caso extremo donde $k=n$ (modelo estimado n veces), cada fold contiene solo una observación para validación y las demás para

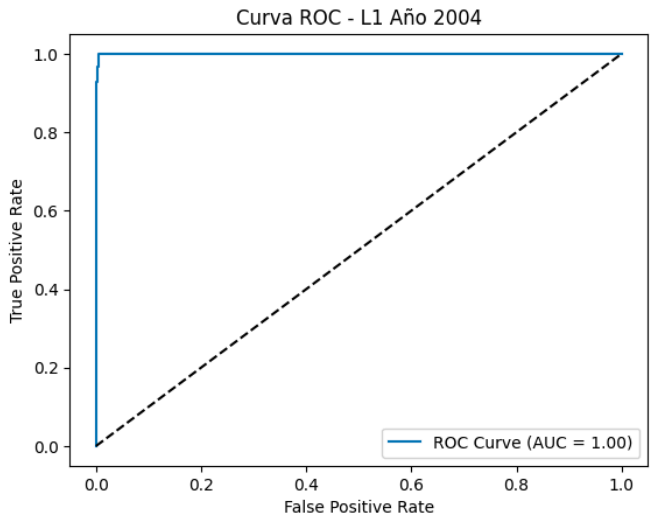
entrenamiento. Este método maximiza el uso de los datos para entrenar el modelo y produce estimaciones del error muy detalladas, pero puede ser más sensible a los valores atípicos, ya que cada observación tiene un peso total en la evaluación del error.

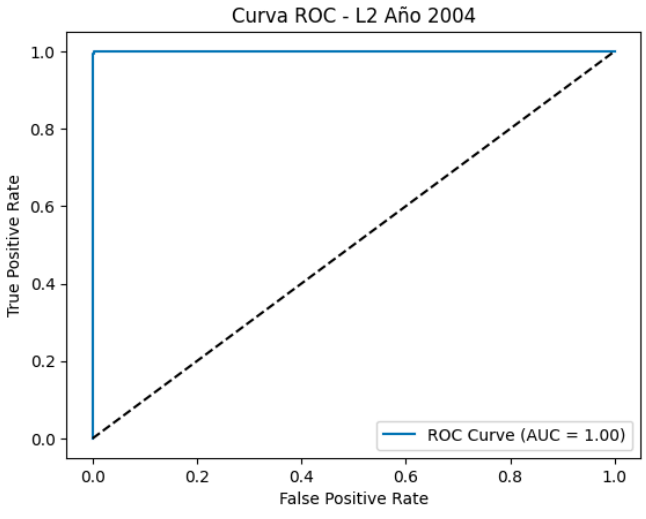
Ejercicio 4

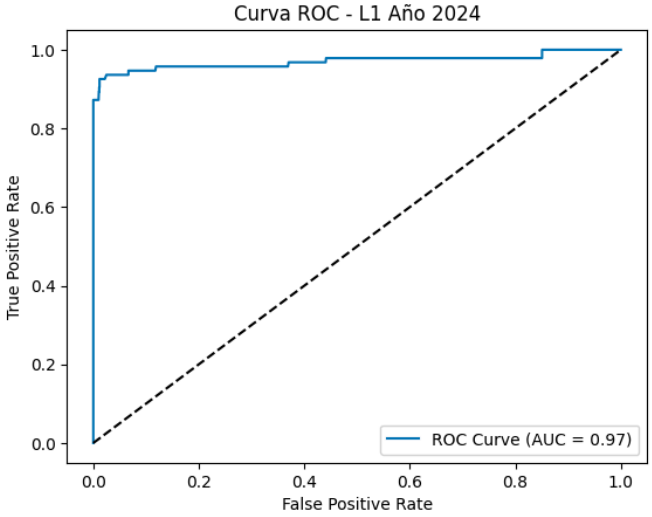
Se implementaron modelos de regresión logística con regularización L1 (LASSO) y L2 (Ridge), con un valor de $\lambda=1$ para ambas penalizaciones. El proceso de entrenamiento se llevó a cabo utilizando los datos de 2004 y 2024 por separado. Antes de entrenar los modelos, se manejaron valores faltantes e infinitos, imputándolos mediante la mediana de cada columna, y se estandarizaron todas las variables predictoras para garantizar estabilidad numérica durante el proceso de optimización. Esto fue especialmente relevante debido a la alta dimensionalidad de los datos, que incluían 5554 variables predictoras en ambos años después de transformar variables categóricas en dummies y garantizar que las bases de datos tuvieran las mismas columnas.

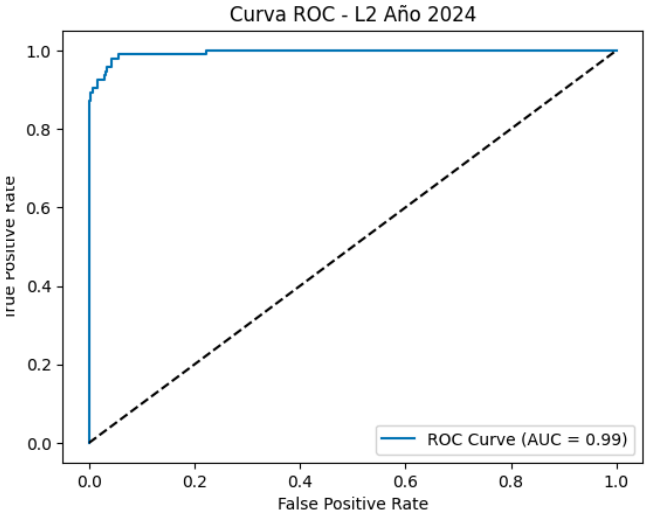
El entrenamiento de los modelos se realizó de forma separada para cada año y penalización. Se utilizaron las métricas de evaluación estándar para modelos de clasificación: matriz de confusión, curva ROC, área bajo la curva (AUC) y precisión (accuracy). Además, se generaron las curvas ROC correspondientes para visualizar el desempeño de los modelos en términos de sensibilidad y especificidad.

Para el año 2004, ambos modelos alcanzaron un AUC de 1.00, indicando una capacidad predictiva perfecta en el conjunto de prueba, con una precisión del 99%. Las matrices de confusión mostraron un número muy bajo de falsos positivos y falsos negativos. En el caso del año 2024, el modelo con LASSO obtuvo un AUC de 0.97, mientras que Ridge logró un AUC de 0.99; ambos alcanzaron una precisión del 99%. Las matrices de confusión en este caso también mostraron diferencias mínimas entre ambas penalizaciones.









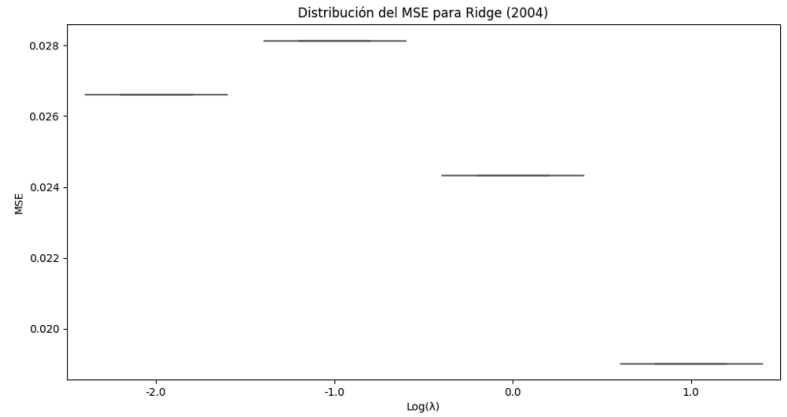
La implementación de regularización mejoró la estabilidad y la capacidad de generalización de los modelos en comparación con el análisis previo realizado en el TP3. En ese ejercicio, los modelos sin regularización presentaron mayor sensibilidad a la alta dimensionalidad y a potenciales problemas de sobreajuste. En este caso, la inclusión de penalizaciones L1 y L2 permitió manejar de manera más eficiente las variables redundantes y ajustar modelos que mantienen un desempeño elevado incluso en conjuntos de prueba.

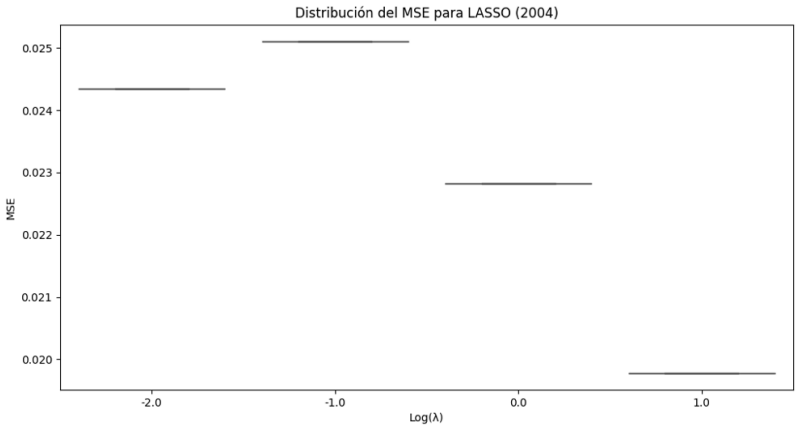
Ejercicio 5

Se analizó el impacto de la regularización mediante Ridge y LASSO en modelos de regresión logística, explorando el comportamiento del MSE y la selección de variables en función de distintos valores del parámetro de regularización λ. Ambos métodos fueron evaluados mediante validación cruzada, comparando la distribución del MSE en cada partición y el efecto sobre los coeficientes del modelo.

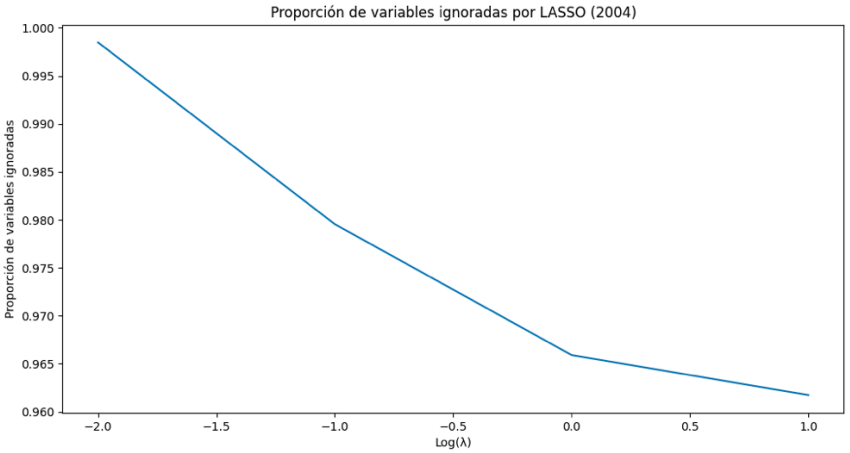
En el caso de Ridge, los boxplots de la distribución del MSE mostraron una tendencia inicial a disminuir conforme λ aumenta desde valores muy pequeños, estabilizandose luego en un rango constante y ligeramente superior a medida que λ crece más allá de un punto óptimo. Esto se debe a que Ridge penaliza el tamaño de los coeficientes, reduciendo su magnitud sin eliminarlos completamente, lo que previene el sobreajuste al mantener todas las variables en el modelo. El λ óptimo seleccionado fue 0.1, donde el modelo logró el mejor equilibrio entre bajo error de predicción y simplicidad estructural. Este comportamiento refuerza la idea de que Ridge es útil para escenarios en los que se prioriza mantener todas las variables con información predictiva, aunque con menor peso. Por otro lado, LASSO

presentó un comportamiento distintivo al combinar la regularización con la selección de variables. Los boxplots del MSE reflejaron estabilidad dentro de cada partición y mostraron una tendencia similar a la de Ridge, con valores mínimos en un rango de λ intermedio.





Sin embargo, el análisis de la proporción de variables ignoradas reveló que, a medida que λ aumenta, el modelo elimina progresivamente más variables al forzar sus coeficientes a ser exactamente cero. Esta capacidad para seleccionar automáticamente un subconjunto de variables relevantes lo hace particularmente útil en contextos donde hay redundancia o ruido en las variables explicativas. Al igual que en Ridge, el λ óptimo para LASSO también fue 0.1, aunque con la ventaja adicional de simplificar el modelo al reducir su dimensionalidad.



De esta manera, los resultados obtenidos evidencian las diferencias clave entre ambos métodos de regularización. Mientras que Ridge es adecuado para problemas donde todas las variables tienen alguna relevancia predictiva, LASSO sobresale en escenarios donde la selección de un subconjunto de variables es deseable. Los gráficos de distribución del MSE y de proporción de variables ignoradas permitieron ilustrar cómo ambas técnicas logran mejorar la capacidad de generalización de los modelos al reducir el riesgo de sobreajuste,

aunque con enfoques distintos. Estos métodos son herramientas esenciales para construir modelos más robustos y fácilmente interpretables en análisis de datos.

Ejercicio 6

En el análisis con LASSO utilizando $\lambda=0.1$, se descartaron 2539 variables, lo que representa aproximadamente el 48% de las 5553 columnas iniciales y el 96% de las 2640 columnas restantes tras el preprocesamiento inicial. Este resultado refuerza la capacidad de LASSO para seleccionar automáticamente las variables más relevantes, eliminando aquellas con menor impacto en la predicción de la desocupación.

Entre las variables descartadas, predominan aquellas con baja varianza, como se anticipó en el inciso anterior, donde se identificaron columnas con valores constantes o distribuciones irregulares como candidatas poco útiles. Estas variables tienen características consistentes con las estadísticas descriptivas obtenidas: valores medios y medianos cercanos a cero, baja varianza y, en algunos casos, patrones extremos o ruidosos. Esto coincide con lo observado en el inciso 1 de la Parte I, donde se destacaron variables como región, aglomerado, ingreso per cápita familiar, y otras relacionadas con la composición del hogar (ocupados y desocupados) como indicadores clave. El modelo LASSO conservó muchas de estas variables relevantes, descartando principalmente aquellas que aportan menor discriminación para predecir la desocupación. Este resultado confirma que el preprocesamiento inicial y el modelo LASSO trabajan de manera complementaria. Antes se identificaron variables con baja variabilidad y potencialmente irrelevantes, mientras que LASSO afianzó esta selección al ajustar coeficientes a cero en aquellas variables cuyo impacto predictivo era limitado. La eliminación de estas características permitió simplificar la dimensionalidad del modelo, manteniendo únicamente aquellas con mayor relevancia para el análisis.

Ejercicio 7

En el análisis se compararon los modelos de regresión logística con regularización LASSO y Ridge en los años 2004 y 2024, evaluando su desempeño en términos del error cuadrático medio (MSE) y la selección de predictores relevantes.

Ridge presentó un buen ajuste en 2004, con MSE promedio entre 0.026 y 0.019, pero su rendimiento empeoró significativamente en 2024, donde los MSE aumentaron a un rango entre 0.068 y 0.112. Esto sugiere que Ridge no se adapta bien a los datos más complejos o

cambiantes de 2024, posiblemente debido a su incapacidad para descartar variables irrelevantes y mitigar el ruido.

Por el contrario, LASSO mostró un mejor desempeño en ambos años. En 2004, el MSE promedio osciló entre 0.024 y 0.019, mientras que en 2024 logró resultados incluso mejores, con MSE entre 0.015 y 0.022. Esto evidencia que LASSO no solo tiene un mejor ajuste a los datos, sino que además se adapta mejor a las complejidades de 2024. Su capacidad para establecer coeficientes en cero y seleccionar únicamente las variables más relevantes es una ventaja significativa frente a Ridge.

En cuanto a la selección de predictores, en 2004 LASSO seleccionó 101 variables de un total de 2640, mientras que en 2024 esta cantidad se redujo a 97. Aunque hubo coincidencias significativas en los predictores seleccionados entre ambos años, algunos predictores considerados relevantes en 2004 no lo fueron en 2024, y viceversa, lo que refleja cambios en los patrones subyacentes y en la importancia relativa de las variables. En contraste, Ridge incluyó todas las variables en el modelo, preservando información completa pero introduciendo ruido que afectó su rendimiento.

De esta manera, LASSO fue el método de regularización más efectivo, logrando un menor MSE en ambos años y adaptándose mejor a las características cambiantes de los datos. Su capacidad para reducir la dimensionalidad al seleccionar únicamente variables relevantes explica su superioridad frente a Ridge, especialmente en 2024, donde la mayor complejidad de los datos hizo aún más evidente esta ventaja.