TRABAJO PRÁCTICO Nº4

CIENCIA DE DATOS

Entrega TP4

Asignatura

Ciencia de Datos

Primavera 2024

Alumnas

Micaela Vollert

Sofía Dillon

Juana Matticoli

Profesores

María Noelia Romero

Ignacio Spiousas

Fecha de entrega

3/11/2024

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

EJERCICIO 1

En la base hogar, identificamos como variables que pueden ser predictivas de la desocupación y podrían perfeccionar el análisis a las relacionadas con la ubicación geográfica, como "región" y "aglomerado", ya que las oportunidades laborales varían según las características económicas de cada región. Además, variables económicas como el "Ingreso Total Familiar (itf)" y los índices de pobreza e ingresos per cápita (ipcf, idecifr, rdecifr) son relevantes, dado que los hogares con ingresos bajos o alta presión económica suelen estar correlacionados con mayores tasas de desocupación. También se podrían considerar variables que reflejan las condiciones de vivienda, como el "tipo de vivienda (iv1)" o el "nivel de hacinamiento (iv6)", ya que estas están relacionadas con la estabilidad económica. Finalmente, la composición del hogar, medida a través del "número total de miembros (IX_Tot)", así como la proporción de personas dependientes (IX_Mayeq10, IX_Men10), puede aportar información valiosa sobre las dinámicas económicas y laborales del hogar. Estas variables juntas pueden proporcionar un panorama integral para modelar la probabilidad de desocupación.

EJERCICIO 2

La consigna solicita integrar información de la Encuesta Permanente de Hogares (EPH) para analizar datos de los años 2004 y 2024. Para ello, se descargan las bases de microdatos de hogares en formatos específicos: .dta para 2004 (Hogar_t104.dta) y .xls para 2024 (usu_hogar_T124.xls). En primer lugar, se carga la base utilizada en el TP3 (base_tp3.csv), que contiene datos individuales, dividiéndola en dos subconjuntos: individual_2004 e individual_2024, según el año. Luego, se cargan las bases de hogares de cada año utilizando las funciones de lectura correspondientes: *read_dta* para 2004 y *read_excel* para 2024. Se filtran ambas bases de hogares para conservar únicamente las observaciones de los aglomerados 32 (CABA) y 33 (GBA), utilizando la función *isin*. Posteriormente, se homogenizan los nombres de las columnas de cada base a minúsculas para facilitar el merge. Las bases de hogares se unen con las de datos individuales de cada año mediante la función *merge*, utilizando como claves las variables CODUSU y NRO_Hogar. Finalmente, las bases resultantes de 2004 y 2024 se combinan en una única base consolidada (df) mediante la función *concat*, la cual se guarda para su posterior análisis.

EJERCICIO 3

El código tiene como objetivo limpiar la base de datos aplicando criterios de manejo de valores faltantes, variables categóricas y valores extremos (outliers). En la primera etapa, se eliminan columnas con un porcentaje significativo de datos faltantes, estableciendo un umbral del 70%; es decir, si una columna tiene menos del 70% de datos válidos, se elimina con la función *dropna*. En la segunda etapa, se procesan las variables en formato string: dado que las variables categóricas no son

compatibles con análisis estadísticos numéricos, se convierten en valores numéricos mediante mapeo, como en el caso de CODUSU, y luego se eliminan las columnas restantes en formato string utilizando *select_dtypes*. En la tercera etapa, los valores faltantes restantes (NaN) se reemplazan con la mediana de cada columna numérica, ya que la mediana es robusta frente a valores extremos. Finalmente, en la cuarta etapa, se eliminan outliers utilizando el rango intercuartílico (IQR): los valores que están fuera de 1.5 veces el IQR por debajo del primer cuartil o por encima del tercer cuartil son descartados. Estas decisiones aseguran que la base final (df3) sea coherente, completa y adecuada para el análisis, reduciendo el ruido estadístico y la influencia de valores extremos.

EJERCICIO 4

Se agregan tres variables nuevas a la base de datos, relevantes para predecir la desocupación. La primera es la proporción de personas ocupadas en el hogar (`ocupados`), calculada agrupando a los individuos por hogar (`CODUSU`) y determinando la proporción de personas con estado laboral ocupado (`ESTADO = 1`). Esta variable es útil porque un mayor número de ocupados podría indicar una menor probabilidad de desocupación en el hogar. La segunda variable es la proporción de personas dependientes en el hogar (`dependientes`), que incluye a quienes son inactivos (`ESTADO = 3`) o menores de 10 años (`ESTADO = 4`). Esta métrica aporta información sobre la carga económica del hogar, ya que una mayor proporción de dependientes podría estar vinculada a un mayor riesgo de desocupación. La tercera variable es el ingreso per cápita familiar (`ingreso_per_capita`), obtenido dividiendo el ingreso total familiar (`itf_y`) por el número total de miembros del hogar (`ix_tot`). Esta variable refleja el nivel de ingresos promedio por persona, lo cual es relevante para entender las condiciones económicas del hogar y su posible relación con la desocupación.

EJERCICIO 5

En este ejercicio se seleccionaron tres variables relevantes de la encuesta de hogar para predecir la desocupación: el ingreso total familiar (`itf_y`), el tamaño del hogar (`ix_tot`) y el nivel educativo promedio del hogar (`nivel_ed`). Posteriormente, se calculan las estadísticas descriptivas básicas (como media, mediana, y percentiles) para estas variables. El análisis revela que los ingresos familiares son muy desiguales, con una mediana de 1.656, pero algunos hogares alcanzan valores extremos cercanos a 350.000. En cuanto al tamaño del hogar, este es relativamente homogéneo, con un promedio cercano a 4 miembros. Por último, el nivel educativo promedio del hogar indica que la mayoría tiene educación secundaria o terciaria incompleta, aunque se observan tanto niveles educativos más bajos (primaria incompleta) como más altos (terciaria o universitaria completa).

PARTE II: Clasificación y Regularización

EJERCICIO 1

Partimos la base de datos ‘df4’ para analizar las observaciones de personas según actividad laboral, basadas en la variable ‘estado’. En primer lugar, realizamos una limpieza de la base, eliminando las filas en las que ‘estado’ tiene valor 0. También notamos que la columna ‘ano4_y’ se encontraba duplicada, por lo que la eliminamos. A continuación, creamos dos bases separadas según los años y en ambas añadimos una nueva variable llamada ‘desocupado’, que se denominó la variable dependiente. El resto de las columnas son variables independientes. Para incluir el intercepto, agregamos una columna de unos en ‘X’. Utilizando el comando *train_test_split*, partimos cada base en conjuntos de ‘prueba’ y de ‘entrenamiento’ en una proporción 70/30, una semilla de 101 para garantizar la reproducibilidad. El tamaño de las bases pueden visualizarse en las **Figuras 1 y 2**.

```
Tamaño de X_train_2004: (2394, 139)
Tamaño de X_test_2004: (1026, 139)
Tamaño de y_train_2004: (2394,)
Tamaño de y_test_2004: (1026,)
```

Figura 1. Tamaño de conjuntos de prueba y de entrenamiento para 2004.

```
Tamaño de X_train_2024: (3941, 139)
Tamaño de X_test_2024: (1689, 139)
Tamaño de y_train_2024: (3941,)
Tamaño de y_test_2024: (1689,)
```

Figura 2. Tamaño de conjuntos de prueba y de entrenamiento para 2024.

EJERCICIO 2

Antes de realizar la regularización, estandarizamos las variables independientes para ambos períodos. La función *standardize_data* usa ‘StandardScaler’ para centrar las variables en cero y escalarlas a una desviación estándar de uno. Luego, realizamos una regresión Ridge con validación cruzada para seleccionar el valor óptimo de λ (o alpha) para los años 2004 y 2024. Hallar el valor óptimo de alpha permitirá ajustar el modelo sin sobreajustarlo al conjunto de entrenamiento, para lograr una mejor generalización. Definimos el rango de valores de alpha. Usamos ‘GridSearchCV’ para realizar una validación cruzada de 5 pliegues (cv=5), probando cada valor de alpha para evaluar cuál proporciona el mejor rendimiento del modelo. Por último, imprimimos el valor óptimo de λ para cada período y en ambos casos nos dio 1.

EJERCICIO 3

En validación cruzada (*cross validation*), el tamaño de *k* tiene efecto directo en la estabilidad y precisión de la estimación del rendimiento del modelo. Usar un *k* muy pequeño implica realizar menos iteraciones de validación, por lo tanto, el modelo está entrenado y evaluado en pocos

subconjuntos. Esto genera una reducción de la capacidad de generalización. Entrenar el modelo con menos subconjuntos implica que las métricas de rendimiento varíen bastante entre las diferentes iteraciones de la validación cruzada, por lo tanto, la varianza será mayor. El sesgo también aumenta porque el modelo se entrena en conjuntos más pequeños, que tienden a subajustar los datos, lo que hace que la estimación de error tenga un sesgo hacia valores más altos. Como el modelo se entrena y se evalúa en pocas particiones, el proceso será más rápido e implica un menor costo computacional. Por otro lado, usar un *k* muy grande permite un análisis más exhaustivo del modelo porque más veces el modelo se entrena con diferentes subconjuntos de datos. Esto mejora la estabilidad de la estimación y aumenta la capacidad de generalización, proporcionando una mejor estimación del rendimiento real y disminuyendo el sesgo. A su vez, al entrenarse en conjuntos muy similares a la totalidad de los datos, el rendimiento del modelo es más consistente entre iteraciones, por lo que también disminuye la varianza. Con un k grande, si los datos son muchos, el modelo será más costoso de entrenar en términos computacionales. En cuanto a *Cross Validation Leave-One-Out* (LOOCV), donde $k=n$, el modelo se entrena n veces, cada vez con un único punto de datos como conjunto de prueba y el resto como conjunto de entrenamiento. En este caso, las estimaciones son muy buenas, casi imparciales. Es útil cuando el conjunto de datos no es muy grande ya que maximiza el uso de esos datos. Por ende, si el conjunto de datos es muy grande, entrenar el modelo n veces llevará mucho tiempo. En LOOCV la estimación puede tener una varianza alta, ya que los conjuntos de datos de entrenamiento no son muy diferentes entre sí, por lo que los resultados pueden ser muy sensibles a las variaciones de los datos individuales.

EJERCICIO 4

Reutilizamos las variables ‘X_train_2004_std’, ‘X_test_2004_std’, ‘X_train_2024_std’ y ‘X_test_2024_std’ previamente estandarizadas en el Ejercicio 2. Creamos dos modelos de regresión logística con penalizaciones L1 (Lasso) y L2 (Ridge) y los entrenamos con el conjunto de entrenamiento estandarizado. El proceso se realizó de la misma manera para ambos períodos. Se evaluaron los modelos utilizando la matriz de confusión, el accuracy, la curva ROC y el área bajo la curva (AUC). Los datos de salida de los dos modelos del año 2004 se ilustran en la **Figura 3 y 4**. Para ambos se reporta un valor AUC de 1, lo que implica que todos los datos fueron clasificados correctamente y los modelos están clasificando correctamente el 100% de las observaciones. No se reportaron falsos positivos ni falsos negativos en la matriz de confusión. Esto indica que los modelos son perfectos en distinguir entre las clases ‘desocupado’ y ‘no desocupado’. Los resultados reportados nos resultan extraños, ya que pareciera que ambos modelos tienen una performance perfecta. Esto puede deberse a que la base de datos extraída del TP3 no es correcta tras arrastrar un error de filtro de la variable ‘estado’.

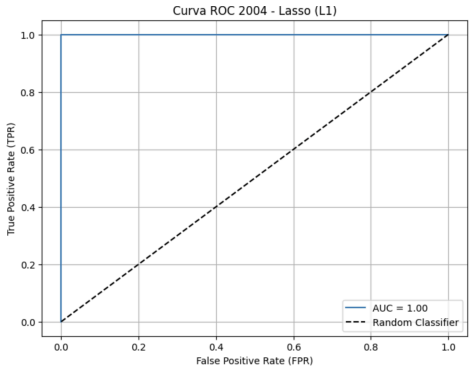

Figura 3. Valor AUC y curva ROC para el año 2004 del modelo LASSO.

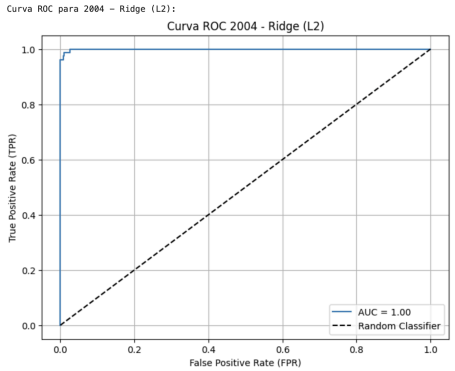

Figura 4. Valor AUC y curva ROC para el año 2004 del modelo Ridge.

Los datos de salida de los dos modelos LASSO y Ridge del año 2024 se ilustran en la **Figura 5 y 6** a continuación. De la misma manera que para 2004, en ambos se reporta un valor AUC de 1 y una curva ROC ideal. No se reportaron falsos positivos ni falsos negativos en la matriz de confusión. Esto refuerza la sospecha de que los datos extraídos del TP3 no son correctos.

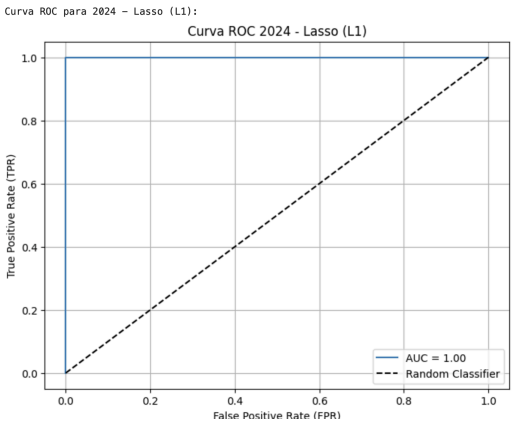

Figura 5. Valor AUC y curva ROC para el año 2024 del modelo LASSO.

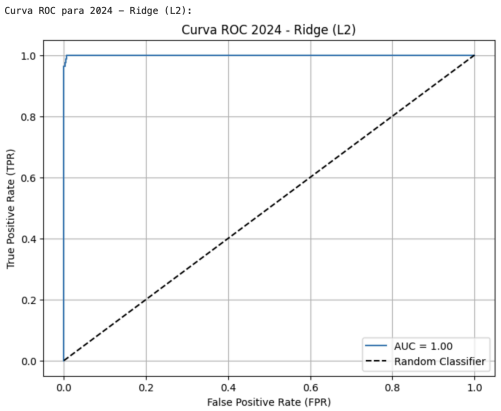

Figura 6. Valor AUC y curva ROC para el año 2024 del modelo Ridge.

En comparación a los modelos de regresión logística del TP3, se evidencia que la *performance* de los modelos de regularización para el año 2024 son superiores ya que reportan mayor accuracy, mayor AUC y una curva ROC ideal. Para el año 2004, tanto en los modelos propuestos en el TP3 como los modelos de regularización reportan valores muy altos, casi perfectos, de AUC y de *accuracy*. Sin embargo, la matriz de confusión reportada en el TP3 demostró que todos las observaciones fueron clasificadas correctamente a excepción de 2 falsos negativos. Por lo tanto, la *performance* de los modelos de regularización son superiores.

EJERCICIO 5

En este ejercicio, se llevó a cabo un barrido de valores de regularización λ para modelos de regresión logística con penalizaciones L1 (LASSO) y L2 (Ridge) con el objetivo de identificar el λ óptimo para cada modelo en los años 2004 y 2024. Dado que en 'scikit-learn' la regularización se controla mediante el parámetro *C*, se definió un rango de *C* equivalente a 10^{n} con *n* entre -5 y 5. Se usó validación cruzada de 10 particiones para evaluar el desempeño de cada valor de *C* y seleccionar el que maximiza ROC-AUC. Los mejores valores de λ obtenidos fueron 0.01 para Ridge y 1.0 para LASSO en ambos años, lo que indica que la regularización óptima para el modelo Ridge es más débil, mientras que para LASSO, una regularización moderada fue suficiente.

En la segunda parte del ejercicio, se realizó un *boxplot* donde se pudo visualizar la distribución del error de predicción (MSE) para cada λ en la regularización Ridge (ver **Figura 7**). Cada *box* corresponde a un valor de λ y contiene la variabilidad del MSE en cada partición de la validación cruzada. La figura ilustra que los valores de $log_{10}(\lambda)$ entre -2 y 1 son los mejores en términos de error de predicción ya que tienen el MSE más bajo y con variabilidad menor. En cambio, para $log_{10}(\lambda) = 4\ y\ 5$ se evidencia que la regularización es demasiado fuerte y que sobreajusta los datos ya que el MSE es notablemente mayor.

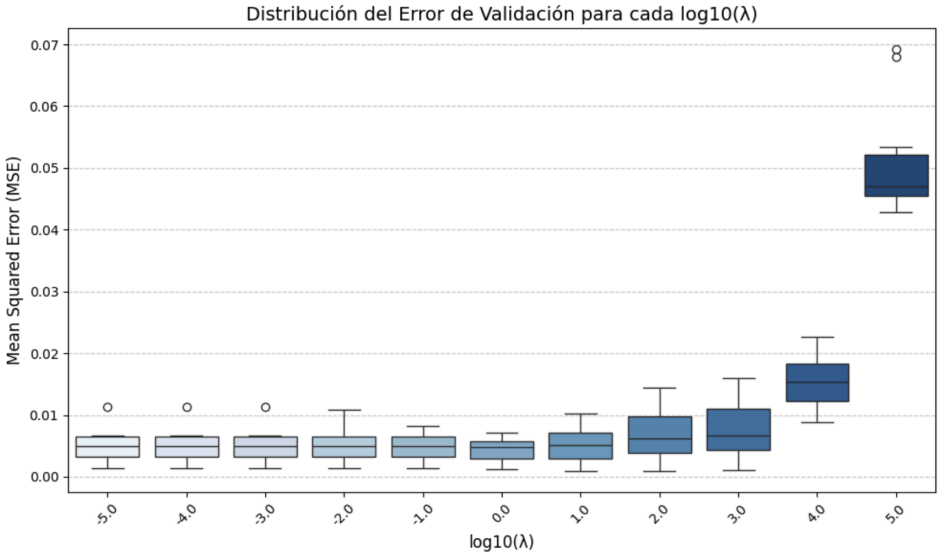

Figura 7. Boxplot de distribución del MSE para cada λ.

Por último, en lo que identificamos como la tercera parte del ejercicio 5, se evaluó la proporción de variables ignoradas en LASSO, graficando cómo el modelo selecciona características en función de λ, ya que a medida que este aumenta LASSO tiende a reducir el número de variables incluidas. La gráfica muestra que valores altos de λ eliminan muchas variables. Esto puede limitar la capacidad predictiva del modelo. En esta misma línea de análisis, un valor de λ demasiado bajo incluirá demasiadas variables, aumentando el riesgo de sobreajuste. Concluimos que un buen valor de λ se aproxima a $log_{10}(\lambda) = -3$ porque en este punto se redujeron significativamente el número de variables pero sigue conservando una proporción considerable para evitar perder información clave.

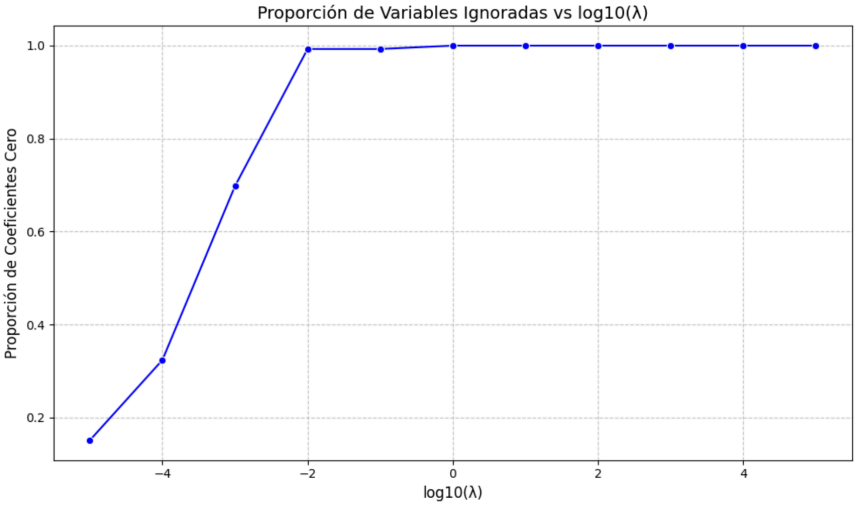

Figura 8. Gráfico de proporción de variables ignoradas vs $log_{10}(\lambda)$.

EJERCICIO 6

En primer lugar, basadas en los resultados obtenidos en el punto anterior, definimos los valores óptimos de lambda para cada año y creamos un diccionario para almacenar los resultados. Para cada período entrenamos un modelo LASSO con su respectivo valor de λ óptimo (0.001 para 2004 y 1.0 para 2024) sobre los datos de entrenamiento normalizados. Luego, obtuvimos los coeficientes de cada variable. Si un coeficiente es igual a cero, la variable será descartada por el modelo ya que LASSO no las considera relevantes.

Hay ciertas variables que inesperadamente fueron descartadas por LASSO para el período de 2004 (en referencia al Ejercicio 1), ya que hubiéramos creído que están relacionadas con la desocupación. Las variables descartadas están asociadas a la ubicación geográfica (‘region_x’, ‘aglomerado_x’, ‘region_y’, ‘aglomerado_y’), a los ingresos (‘itf_x’, ‘itf_y’, ‘ipcf_x’, ‘ipcf_y’, ‘decifr_x’, ‘rdecifr_x’, ‘gdecifr_x’) y a condiciones de la vivienda (‘iv1’, ‘iv6’, ‘ix_tot’, ‘ix_men10’, ‘ix_mayeq10’). Sin embargo, algunas variables relacionadas con la composición del hogar y demografía sí fueron incluídas por el modelo como ‘nro_hogar, ‘ch03’, ‘nivel_ed’, ‘estado’, las cuales contemplan a características individuales y familiares. Una posible explicación para el descarte de estas variables podría ser la existencia de multicolinealidad con otras ya seleccionadas o podría ser que no tienen suficiente variabilidad para explicar diferencias en la desocupación en 2004.

En cuanto al modelo del período del 2024, nos resultó extraño que todas las variables que fueron seleccionadas como relevantes en el Ejercicio 1 fueron descartadas por LASSO. Entre ellas se encuentran: ‘región’, ‘aglomerado’, ‘itf’, ‘ipcf’, ‘iv1’, ‘iv6’, etc. El descarte de la totalidad de las variables implica que no se detectó una relación estadísticamente significativa entre los predictores y la desocupación en los datos de 2024. Esto puede deberse a problemas con los datos utilizados, por ejemplo desbalance de clases o alta correlación entre las variables, o a un ajuste inadecuado del modelo.

EJERCICIO 7

Utilizando los valores óptimos de regularización obtenidos previamente, comparamos los modelos de regresión logística Ridge y LASSO para los años 2004 y 2024. Para ello, para cada combinación de modelo y año, entrenamos un modelo Ridge (penalización L2) o LASSO (penalización L1) sobre los datos de entrenamiento normalizados y generamos predicciones para el conjunto de prueba. Calculamos el error cuadrático medio (MSE) entre las predicciones y los valores reales de prueba para evaluar el desempeño.

Los MSE reportados para ambos modelos del 2004 son muy bajos: 0.00292 para Ridge, 0.00194 para LASSO. Suponemos que esto sugiere que la regularización no tuvo un impacto significativo en la capacidad predictiva de los modelos en este conjunto de datos. El rendimiento de Ridge y LASSO fue similar, lo que podría indicar una relación bastante evidente entre las variables regresoras y la desocupación. También puede significar que las variables ya estaban bien

seleccionadas y no necesitaban una fuerte selección de características por parte de LASSO. Por otro lado, reportamos un MSE de 0.00059 para el modelo de Ridge del 2024 y un MSE de 0.0 para el modelo de LASSO de 2024. El modelo de regularización Ridge está ajustando de manera adecuada, capturando la relación entre las variables y la desocupación, pero hay un pequeño error residual. El modelo de regularización LASSO ha demostrado una predicción perfecta en el conjunto de *test*, pero esto se debe probablemente a que descartó todas las variables predictoras o encontró un conjunto de variables muy restringido que proporciona una predicción casi perfecta, que puede ser indicador de un sobreajuste (*overfitting*).

Podemos concluir que los modelos que demostraron un peor desempeño fueron los del año 2004, tanto Ridge como LASSO. Sin embargo, el valor del MSE es muy bajo, lo cual indica que tienen una muy buena capacidad predictiva. En cuanto a los modelos predictivos para el 2024, creemos que LASSO realizó un descarte muy agresivo y estricto de variables. Por lo tanto, Ridge mantiene un MSE muy bajo, lo cual indica que capturó alguna relación más sutil entre las variables en 2024 que LASSO no. Se sugiere que mantener las variables correlacionadas puede ser una mejor estrategia. Ridge podría considerarse más robusto y generalizable en 2024.

Por último, observamos que LASSO hizo una selección distinta de predictores en 2004 que en 2024. En 2004, LASSO seleccionó un conjunto reducido de variables, indicando que había relaciones significativas con la desocupación. En 2024, LASSO descartó todas las variables o dejó muy pocas. Las causas de estas diferencias pueden ser:

I. Hubo cambios estructurales en los factores socioeconómicos entre 2004 y 2024, haciendo que las variables que antes eran relevantes para predecir la desocupación ya no lo sean.

II. Problemas con la base de datos de 2024: alta colinealidad entre variables pudo haber generado que LASSO descarte todas, poca variabilidad en cada variable o desbalance entre la cantidad de observaciones de las categorías de ‘estado’.