Trabajo práctico N° 4

Informe escrito

Autoras

Clara Bernhardt

María Teresa Laszeski

Rosario Luque

Profesores

María Noelia Castillo

Ignacio Spiousas

Asignatura

Ciencia de Datos

Semestre y año de presentación

2° semestre 2024

Parte I: Analizando la base

1. Diseño de registros de la base Hogar

A partir de la exploración del diseño de registro de la base de hogar, las variables que se consideran potencialmente predictivas de la desocupación son: 1) el ITF (Ingreso Total Familiar) puede reflejar la estabilidad económica del entorno familiar; 2) el IPCF (Ingreso Per Cápita Familiar) ajusta el ingreso familiar según el tamaño del hogar, por lo que puede ser un predictor más preciso del bienestar económico por persona; 3) IX_Tot (Cantidad de miembros del hogar) y IX_Men10 (Cantidad de menores de 10 años), dado que hogares más grandes o con mayor número de dependientes pueden enfrentar mayores cargas económicas, lo que puede influir en las tasas de desempleo de los adultosy 4) II7 (Régimen de tenencia del hogar), dado que los hogares que no son propietarios podrían enfrentar más inestabilidad económica, lo que puede relacionarse con mayores niveles de desocupación.

2. Pre procesamiento de datos

En el proceso de limpieza y preparación de datos, se abordó la tarea de unificar y depurar tanto las bases de Individuos como de Hogares de la Encuesta Permanente de Hogares (EPH) correspondientes al primer trimestre de 2004 y 2024. Inicialmente, se descargaron los archivos en formatos .dta y .xls, respectivamente, desde la página web del INDEC, para luego cargarlos en Visual Studio Code. Con el fin de facilitar la convergencia entre las cuatro bases, se procedió a convertir todos los nombres de variables a minúsculas. Posteriormente, se aplicó un filtro para retener únicamente los datos correspondientes a los aglomerados de Ciudad Autónoma de Buenos Aires (32) y Gran Buenos Aires (33).

En lo que respecta a las bases de individuos, se tomó la decisión de retener la mayor cantidad de variables posibles de la sección “Características de los miembros del hogar”. En este sentido, se tuvo en cuenta la cantidad de *missing values* y la cantidad de respuestas del tipo “Ns./Nr.”, lo cual se consideró como un *missing value* encubierto, de forma tal que aquellas variables que causaban la pérdida de información pronunciadamente fueron descartadas de entrada. Además se incluyeron las variables de ingresos no laborales (menos t_vi) e ipcf (ingreso per cápita).

Para las bases de hogares se siguieron los mismos criterios de selección nombrados anteriormente, lo cual llevó a la elección minuciosa de las variables de la sección “Características de la vivienda” y de las de “Estrategias del hogar”. Además, se tomaron las variables de ix_tot (cantidad de miembros del hogar) e ix_mayeq10 (cantidad de miembros del hogar de 10 y más años), sumado a itf (ingreso total familiar).

A continuación, se procedió a transformar las variables categóricas del 2004 al formato del 2024, es decir, numérico, en vistas de la transformación posterior a dummies que sería realizada, tanto para la base de hogares como la de individuos. Luego, se eliminaron los *missing values* de las variables de interés, que por cuestiones de espacio no se explicará el detalle de la limpieza pero se invita al lector a revisar el código comentado de respaldo. Por último, se revisó que las variables numéricas referentes a ingresos y la variable de edad (ch06) no tuvieran valores negativos y, en caso de poseerlos, estos fueron eliminados para no contaminar los datos.

Una vez eliminados los valores faltantes y los valores sin sentido, además de transformar las variables, se procedió a unir las cuatro bases de datos, para lo cual fueron usadas las variables CODUSU y NRO_Hogar para realizar el *merge*. A continuación, se utilizó el método de la Desviación Absoluta de la Mediana (MAD) para identificar outliers debido a su robustez (Rousseeuw, 1990), fijando un umbral de 3.5 y etiquetando las filas sin outliers con un 0 y las que tenían outliers con un 1. Finalmente, se creó una nueva base de datos con todas las filas etiquetadas con 0, lo cual resultó en un conjunto de 12543 datos depurados y 65 variables. Además, se crearon las variables de PEA

(Población Económicamente Activa), PET (Población en Edad para Trabajar, personas entre 15 y 65 años) y Desocupado, la variable a ser predecida, la cual toma el valor 1 si la persona está ocupada y 0 sino. Asimismo, a partir de la base de hogares se crearon las variables de proporción de ocupados (PEA/ix_tot), proporción que reciben subsidios (v5/ix_tot) y proporción de mayores de 10 (ix_mayeq10/ix_tot). Finalmente, todas las variables categóricas que tomaran dos valores fueron transformadas en binarias y se crearon dummies a partir de aquellas que fueren categóricas. Esto llevó a la creación de la base final que se utilizará luego para el análisis predictivo del desempleo, con un total de 151 variables depuradas.

3. Estadística descriptiva de variables de interés

La variable iv4, que representa el tipo de cubierta exterior del techo de un hogar, podría ser un indicador relevante para predecir el desempleo debido a su potencial para reflejar el nivel socioeconómico y la calidad de vida de los habitantes. Los datos muestran una distribución variada de tipos de techos tanto en 2004 como en 2024 (**Figura 1**), con cambios notables en algunas categorías. Si bien se observa una disminución en techos de membranas/cubiertas asfálticas (2331 en 2004 contra 1254 en 2023) hay una tendencia positiva hacia techos de baldosas/losas (1845 en 2004 contra 2115 en 2024). Esta evolución podría indicar mejoras en las condiciones de vivienda a lo largo del tiempo.

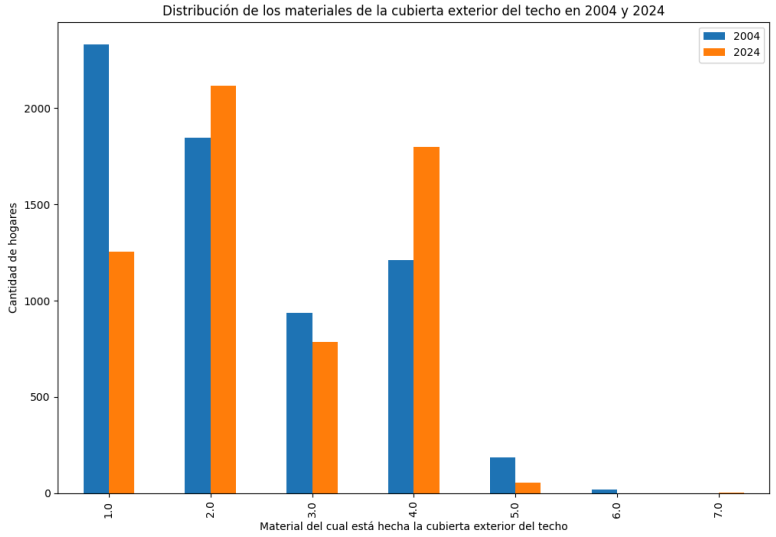

Figura 1. Histograma de la distribución de los materiales de la cubierta exterior del techo en los años 2004 y 2024.

Por su parte, la variable iv7, que refleja las fuentes de agua en los hogares, refleja una distribución interesante entre el 2004 y 2024 (**Figura 2**), que si bien vio una ligera disminución en el acceso a la red pública de agua (4904 en 2004 contra 2836 en 2024) es aún más notable su reducción en el uso de perforaciones con bombas a motor (1561 en 2004 contra 1146 en 2024). Analizar la evolución de las fuentes de agua puede proporcionar *insights* valiosos sobre el desarrollo económico de diferentes áreas y, ser una variable potencialmente predictiva del desempleo.

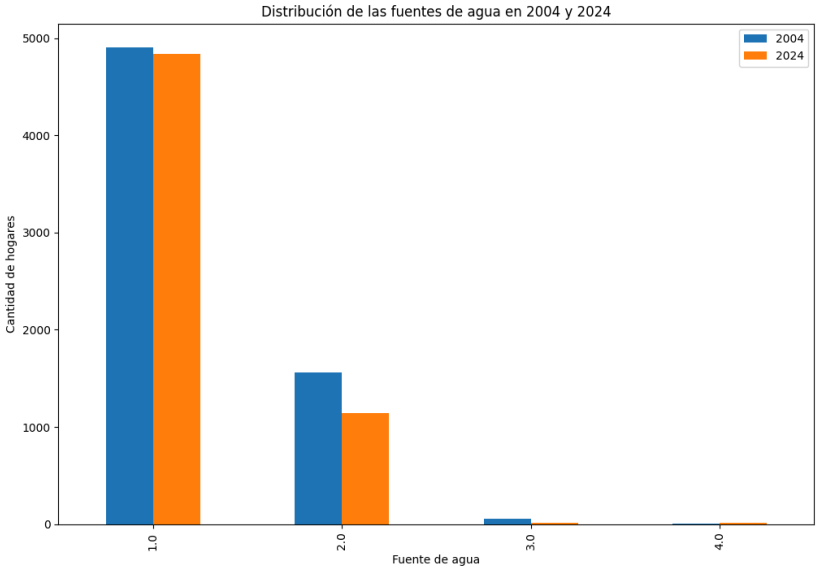

Figura 2. Histograma de la distribución de las fuentes de agua en los años 2004 y 2024.

Finalmente, la variable iv9, que indica la ubicación del baño o letrina en los hogares, refleja hallazgos interesantes (**Figura 3**). En 2004, 6007 hogares tenían el baño dentro de la vivienda, 517 fuera pero dentro del terreno, y 5 fuera del terreno. Para 2024, se observa una ligera disminución a 5883 hogares con baño dentro de la vivienda, una reducción importante a solo 115 hogares con baño fuera pero dentro del terreno, y la aparición de 16 casos con baño fuera del terreno. La ubicación del baño puede ser un indicador de la calidad de la vivienda, el acceso a servicios básicos y el nivel socioeconómico general, factores estrechamente relacionados con las oportunidades de empleo y la estabilidad laboral.

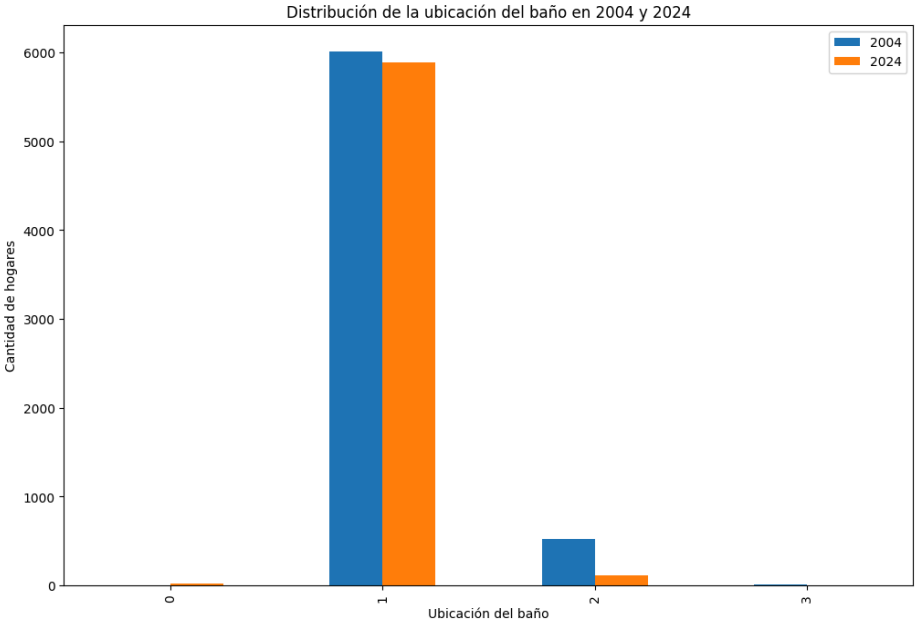

Figura 3. Histograma de la distribución de la ubicación del baño en los años 2004 y 2024.

Parte II: Clasificación

Una vez realizado el preprocesamiento de datos y agregadas las columnas indicadas, se procedió a entrenar y evaluar distintos modelos con el objetivo de intentar predecir si una persona está desocupada o no. El proceso comenzó con la preparación de los datos, utilizando el comando

train_test_split para dividir cada conjunto anual en muestras de entrenamiento (70%) y prueba (30%), con una semilla aleatoria de 101. Se estableció "desocupado" como la variable dependiente, mientras que las demás variables se codificaron como predictores independientes, añadiendo una columna de unos para el término de intercepción.

Aunque no se implementó explícitamente la selección de λ (alpha) mediante validación cruzada, se discutió su importancia teórica. El proceso implicaría dividir el conjunto de entrenamiento en subconjuntos, probar diferentes valores de λ, y seleccionar el que proporcione el mejor rendimiento promedio, evitando el uso del conjunto de prueba para prevenir sesgos y sobreajuste. Por otro lado, las implicancias de elegir diferentes valores de k en la validación cruzada, señalando que un k pequeño puede llevar a estimaciones de error con alta varianza pero bajo sesgo computacional, mientras que un k grande reduce la varianza pero aumenta el sesgo y el costo computacional.

Se llevó a cabo un análisis de regresión logística con regularización para predecir el desempleo en los años 2004 y 2024. Se implementaron dos tipos de regularización: Ridge (L2) y LASSO (L1), ambas con un parámetro λ fijado en 1. Los resultados mostraron diferencias significativas entre los métodos de regularización y los años analizados. Los modelos Ridge demostraron un rendimiento notablemente superior en comparación con los modelos LASSO. Para el año 2004, el modelo Ridge alcanzó una precisión del 92.65% con un MSE de 0.0735, mientras que en 2024, la precisión aumentó al 95.35% con un MSE de 0.0465. Los modelos LASSO, aunque con precisiones similares (92.70% en 2004 y 95.29% en 2024), mostraron un rendimiento inferior en términos de capacidad discriminativa.

Las matrices de confusión proporcionaron insights adicionales sobre el rendimiento de los modelos. Todos ellos exhibieron un alto número de verdaderos negativos, indicando una fuerte capacidad para identificar correctamente los casos de no desempleo. Sin embargo, se observó una tendencia a subestimar los casos de desempleo, especialmente en los modelos de 2024 **(Figuras 4 y 5**).

Ridge					
2004			2024		
	Positivo	Negativo	Positivo	Negativo	
Positivo	24	15	74	1711	Negativo
Negativo	129	1791	10	10	Positivo

Figura 4. Matriz de confusión con Ridge para 2004 y 2024.

LASSO					
2004			2024		
	Positivo	Negativo	Positivo	Negativo	
Positivo	23	13	76	1712	Negativo
Negativo	130	1793	8	9	Positivo

Figura 5. Matriz de confusión con Lasso para 2004 y 2024.

El análisis de los valores de AUC marcó aún más la superioridad de los modelos con regulación Ridge. Con valores de AUC de 0.8956 para 2004 y 0.8891 para 2024, los modelos Ridge demostraron una excelente capacidad discriminativa. En contraste, los modelos LASSO mostraron un rendimiento cercano al azar, con valores de AUC de 0.5543 y 0.4967 para 2004 y 2024,

respectivamente. Por otro lado, las curvas ROC revelan una clara superioridad de los modelos Ridge sobre los LASSO en ambos años. Las curvas de Ridge se sitúan significativamente por encima del clasificador aleatorio y muestran un rendimiento consistente entre 2004 y 2024. En contraste, las curvas de los modelos LASSO se acercan mucho a la diagonal del clasificador aleatorio, lo que sugiere que la penalización L1 puede haber sido demasiado severa, eliminando características predictivas importantes. (**Figura 6**).

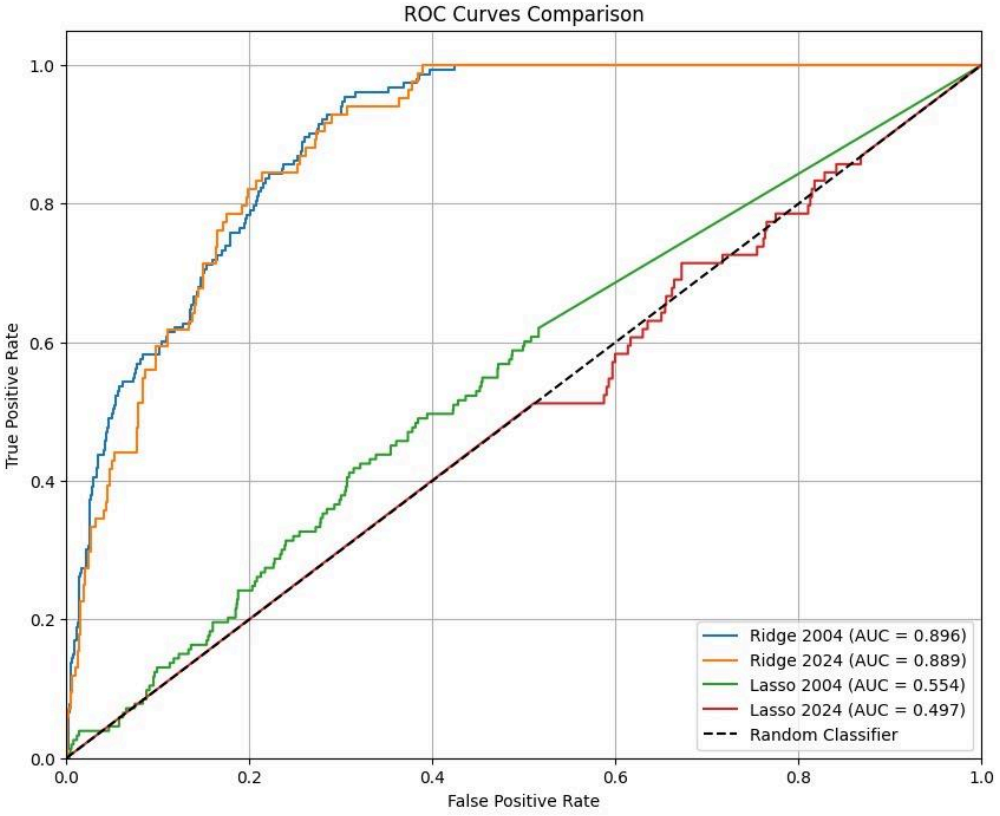

Figura 6. Curvas ROC de los modelos regularizados por Ridge y Lasso para 2004 y 2024. La línea punteada roja representa la línea de azar.

Al comparar estos resultados con los obtenidos en el trabajo práctico 3, se observa que los modelos con Ridge mantienen un rendimiento competitivo e incluso superan ligeramente los resultados del trabajo anterior. Para 2004, aunque la precisión es ligeramente inferior al 0.9325 del TP3, el AUC de nuestro modelo Ridge (0.8956) supera al 0.8779 anterior. En el caso de 2024, nuestro modelo Ridge supera tanto en precisión (0.9535 vs 0.9521) como en AUC (0.8891 vs 0.8709) al mejor modelo del TP3, que era LDA.

A continuación, se analizaron dos métodos de regularización en regresión logística: Ridge (penalización L2) y LASSO (penalización L1). Se utilizó validación cruzada (10-fold CV) para identificar el valor óptimo de regularización (λ), explorando un rango definido como $\lambda = 10^n$, con n $\in \{-5,-4,-3,\ldots,+4,+5\}$. El objetivo principal fue comparar los modelos en términos de precisión predictiva y analizar el impacto de la regularización LASSO sobre la selección de variables. Se generaron boxplots para comparar la distribución del error de predicción para cada valor de λ. Para LASSO, se analizó la proporción de coeficientes nulos ($\beta_i = 0$) en función de λ.

Los resultados reflejan, por un lado, la variación en la precisión de predicción para diferentes valores de λ en Ridge y LASSO. En el caso de Ridge (**Figura 7**), para ambos años (2004 y 2024), la precisión tiende a estabilizarse para λ moderados (alrededor de 10), mientras que los valores extremos de λ generan mayor dispersión y error. En el caso de LASSO (**Figura 8**), el comportamiento es

similar al de Ridge, sin embargo, hay una mayor sensibilidad a valores altos de λ, con errores más pronunciados. Por otro lado, los gráficos de **Figura 9** y **Figura 10** destacan cómo aumenta la proporción de variables ignoradas a medida que λ crece, para ambos años, el modelo elimina progresivamente más variables con regularización más fuerte, alcanzando casi un 100% de exclusión para valores altos de λ. Por último, los valores óptimos de λ para Ridge son: 2004: λ=1.0, 2024: λ=10.0; y para LASSO son: 2004 y 2024: λ= 10.0.

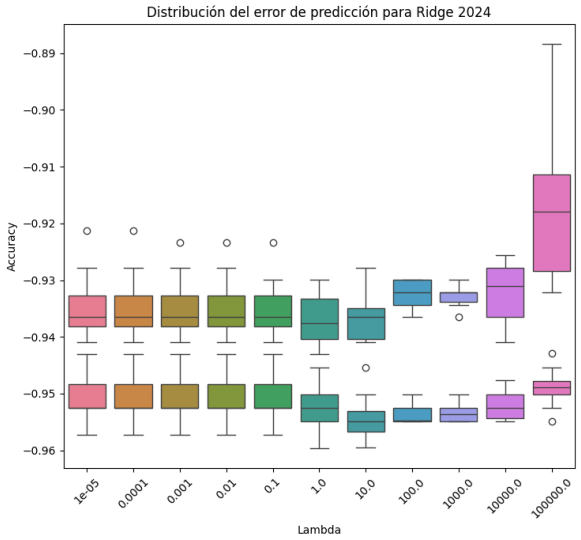

Figura 7. Distribución del error de predicción para Ridge 2024.

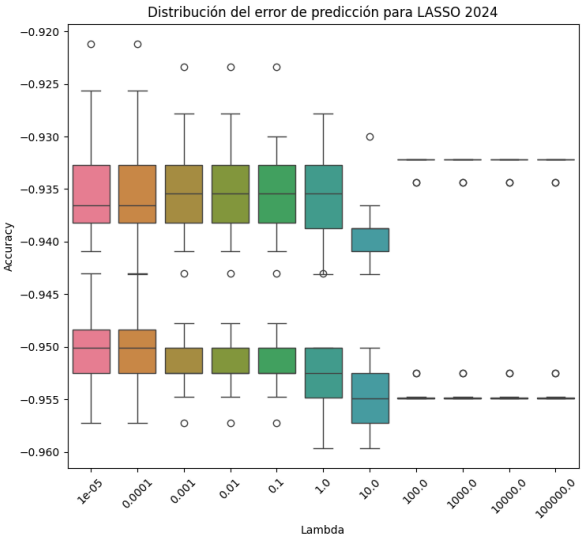

Figura 8. Distribución del error de predicción para LASSO 2024.

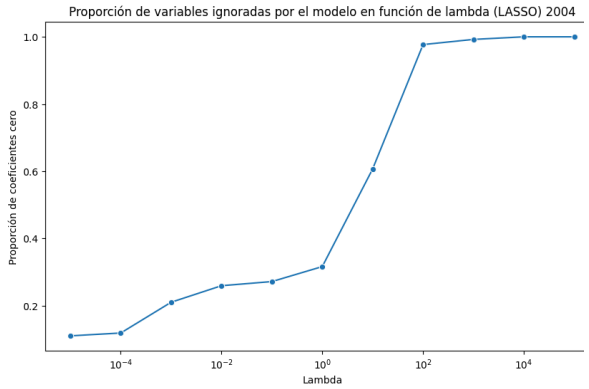

Figura 9. Proporción de variables ignoradas por el modelo en función de lambda (LASSO) 2024.

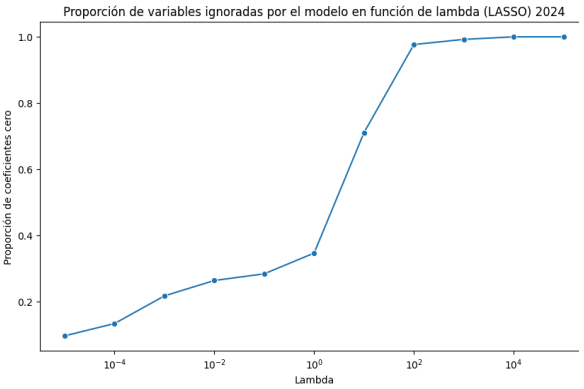

Figura 10. Proporción de variables ignoradas por el modelo de función de lambda (LASSO) 2024.

En conclusión, ambos métodos presentan un rendimiento consistente al optimizar λ para los datasets de 2004 y 2024. Ridge mostró una menor sensibilidad al cambio de λ mientras que LASSO fue más afectado por valores extremos, dada su naturaleza de selección de variables. En cuanto al efecto de la regularización de LASSO, el análisis de la proporción de variables ignoradas confirma que LASSO promueve un modelo más esparzo. Este comportamiento es especialmente útil en contextos donde la interpretabilidad y la simplificación del modelo son importantes. Finalmente, si comparamos Ridge y LASSO, Ridge es preferible en escenarios donde todas las variables deben

mantenerse, ya que minimiza coeficientes sin eliminarlos. En cambio, LASSO es ideal para escenarios de alta dimensionalidad, donde ciertas variables pueden ser irrelevantes o ruidosas.

El análisis hasta ahora realizado revela que el modelo LASSO destaca una lista considerable de variables tanto en la base de datos de 2004 como en la de 2024. Por un lado, las variables destacadas por LASSO de la base de 2004 son "ix_tot" (tamaño del hogar), "prop_ocupados", "prop_subsidio", "prop_mayores_10" y varias dummies relacionadas con características de la vivienda, ingresos y nivel educativo fueron descartadas. Por otro lado, las variables destacadas por LASSO de la base de 2024 además de las variables similares a la base anterior, son "itf" (Ingreso Total Familiar), y varias relacionadas con características individuales, como dummies educativas ("nivel_ed") y variables de composición del hogar.

En el inciso inicial, en la Parte I, se plantearon como variables relevantes, variables relacionadas con la estructura económica del hogar ("itf", "ipcf", y "prop_subsidio"), composición del hogar ("ix_tot" y "ix_men10") y régimen de tenencia y calidad de vivienda ("iv7", "iv9"). Sin embargo, el modelo LASSO ha descartado muchas de estas variables. Esto puede deberse a: multicolinealidad, baja relevancia predictiva o dummies excesivas. Se podría decir que es coherente con lo esperado, ya que la eliminación de variables redundantes o de bajo impacto es esperada en modelos de regularización como LASSO. Si las variables descartadas tienen alta correlación con otras que permanecen, el modelo prioriza mantener solo una. Pero, en parte no, porque resulta curioso que variables como "itf", "ix_tot" y "prop_ocupados", que tienen una conexión teórica clara con la desocupación, hayan sido descartadas. Esto podría indicar que estas variables están capturadas indirectamente por otras (como proporciones relacionadas con la PEA o ingresos per cápita), como también que hay ruido en los datos o no son tan predictivas como se pensaba inicialmente.

En resumen, la relación con la Parte I es indirecta. Si bien las variables propuestas inicialmente reflejan una lógica sólida, el modelo LASSO nos lleva a un análisis más centrado en el poder predictivo empírico, dejando de lado aquellas que no contribuyen mucho a predecir la desocupación.

En el análisis comparativo entre los modelos de regularización Ridge y LASSO, los resultados obtenidos a través de las métricas de precisión, MSE, AUC y las curvas ROC indican una clara superioridad de Ridge sobre LASSO en ambos años. Para el modelo Ridge, se observó un rendimiento consistente en ambos períodos: en 2004, la precisión fue de 0.9265, el MSE fue 0.0735 y el AUC alcanzó 0.896; mientras que en 2024, la precisión aumentó a 0.9551, el MSE disminuyó a 0.0449 y el AUC fue de 0.889. Estos resultados reflejan una alta capacidad discriminativa y robustez en la captación de patrones en los datos a lo largo del tiempo. En cambio, los modelos LASSO mostraron un rendimiento inferior. En 2004, LASSO alcanzó una precisión de 0.9255, un MSE de 0.0745 y un AUC de 0.554, mientras que en 2024, la precisión es de 0.9562, un MSE de 0.0438, pero el AUC cayó, a 0.497. Este bajo en el AUC sugiere que la regularización L1 fue demasiado severa, eliminando características predictivas cruciales, lo que resultó en un modelo con capacidad de predicción muy limitada.

Las curvas ROC de la **Figura 6** corroboran estos hallazgos, ya que las curvas para el modelo Ridge se mantuvieron bien por encima de la línea de azar en ambos años, evidenciando un buen rendimiento discriminativo. En cambio, las curvas para LASSO estuvieron muy cerca de la diagonal del clasificador aleatorio, indicando una deficiente capacidad para separar las clases correctamente. Al comparar estos resultados con los obtenidos en el trabajo práctico 3, se observa que el modelo Ridge ha mejorado en comparación con los resultados previos: en 2004, el AUC de Ridge (0.896) supera al del TP3 (0.8779), y en 2024, el AUC de Ridge (0.889) también supera al AUC del modelo más

preciso del TP3 (LDA, con 0.8709). La precisión de Ridge también fue superior en 2024, pasando de 0.9535 en el TP3 a 0.9551 en este análisis.

En conclusión, Ridge se mostró como el método de regularización más efectivo en este análisis, manteniendo un buen rendimiento a lo largo del tiempo sin sacrificar características importantes, a diferencia de LASSO, que eliminó demasiadas variables predictivas relevantes. Las diferencias observadas entre 2004 y 2024 fueron menores para Ridge, mientras que LASSO presentó un rendimiento notablemente peor en 2024, lo que sugiere problemas de generalización. Por lo tanto, Ridge es la opción más robusta y confiable para este tipo de análisis.

Referencias

INDEC. (2024). Mercado de trabajo. Tasas e indicadores socioeconómicos (EPH) (Trabajo e Ingresos Vol. 8, N° 7). Recuperado de https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-31-58

Rousseeuw, P. J. (1990). Robust estimation and identifying outliers. Handbook of statistical methods for engineers and scientists, 16, 16-11.