Universidad de
SanAndrés

Trabajo Práctico 4

Ciencias de Datos

27 de noviembre 2024

Profesor: **Ignacio Spiousas**

Zahira Chaia

Martina Lucas

PARTE 1: Análisis de la base de hogares y tipo de ocupación

1. En primer lugar, sería importante retomar la definición de desocupación según el INDEC. La población desocupada se refiere a las personas que desean trabajar, buscan empleo activamente, pero no logran conseguirlo. No incluiríamos otras formas de precariedad laboral, como empleos transitorios, trabajos con pocas horas, quienes han dejado de buscar empleo por falta de oportunidades, empleos con salario mínimo o puestos por debajo de su nivel de calificación.

En la base hogar hay múltiples variables indicadoras de la situación socioeconómica del hogar, que podrían ser predictivas de la desocupación. Entre las más representativas estarían “v5” y “v6”, variables binomiales que indicarían si las personas del hogar reciben algún tipo de subsidio o ayuda social en forma de dinero, mercadería, ropa o alimentos. Estas variables aportarían información valiosa, ya que, si el grupo familiar recibe ayuda de este tipo, probablemente se encontraría en situación de necesidad.

Además, podríamos incluir la variable “v12”, que señalaría si el hogar recibió cuotas de alimentos o ayuda en dinero de personas externas al hogar. Esta variable sería indicadora de la situación económica de la familia y de la dificultad de conseguir ingresos mediante una ocupación formal. Por otro lado, consideraríamos la variable “v15”, que refiere a la obtención de préstamos por parte de bancos o financieras. Esta variable sería relevante, ya que, en la mayoría de los casos, para recibir un préstamo de una institución financiera sería necesario demostrar capacidad de pago futura. Finalmente, incluiríamos la variable "v19_a", que evaluaría si menores de 10 años del hogar ayudaron con dinero trabajando en los últimos tres meses.

Por otra parte, ciertas variables que describen las características del hogar serían grandes indicadores del ingreso de sus habitantes y, por ende, podrían estar relacionadas con la desocupación. Entre ellas, consideraríamos “IX_TOT” (cantidad de miembros del hogar) e “ITF” (monto de ingreso total familiar), que aportarían información útil para análisis posteriores. Con esta selección de variables, construiríamos un nuevo dataset enfocado en el análisis de la desocupación.

2. Posteriormente, eliminamos todas las observaciones que no correspondan a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires, y unimos las bases “individual” y “hogar” en una sola base de datos realizando un *merge()* utilizando las

variables **CODUSU** y **NRO_Hogar** como claves para la combinación. Decidimos no unir las bases de ambos años ya que no era necesario para los análisis posteriores, además de que evitaba posibles problemas relacionados a los nombres de las variables.

3. Para realizar un análisis más enfocado y efectivo, decidimos seleccionar algunas columnas relevantes. Las columnas seleccionadas finalmente fueron: ['ch04', 'ch06', 'ch07', 'ch08', 'ch09', 'nivel_ed', 'estado', 'cat_inac', "v5","v6","v12","v15","ix_tot", "iv1","iv2","iv5","iv6", "iv8", "iv12_3", "ii1", "codusu"]. Esta selección nos permitió trabajar con un conjunto de datos más específico, enfocado en las variables que podían relacionarse con la predicción de la desocupación.

Lo primero que hicimos fue evaluar si había datos duplicados y los eliminados. Además, identificamos los datos faltantes en la base con la función *isna*() y observamos que no había datos sin completar. Por otro lado, evaluamos mediante una función en qué filas había valores Ns./Nr o 9 (que equivale a Ns./Nr en la base de 2024). Como había muy pocos valores decidimos eliminarlos, ya que no aportaban datos significativos para el análisis (Diligent, n.d.).

A continuación, evaluamos las variables numéricas, comenzando con la variable "ch06" (Edad). Para ello, elaboramos un histograma que nos permitió visualizar la distribución de la población según su edad. Durante este proceso, también analizamos la presencia de registros negativos, encontrando 50 casos que representaban aproximadamente el 0.71 del total de los datos. Debido a que estos valores no eran coherentes con el análisis, decidimos eliminarlos. Luego hicimos un procedimiento similar con las variable “ix_tot” (cantidad de personas por hogar) y “ii1” (cantidad de habitaciones). Durante este proceso, también analizamos la presencia de registros negativos y no encontramos ninguno.

A continuación evaluamos las variables que tenían solo dos categorías y las renombramos para que sus valores sean 0 y 1. Comenzamos eliminamos en "ch09" la categría "Menor de 2 años", reemplazandola por “no” (ya que estos no saben leer). Luego, con el objetivo de que las variables binarias se convieran en variables dummy. En la base de 2004 para esto utilizamos una función que reemplaza “si” por 1 y “no” por 0 y en la ase de 2024 para esto utilizamos una función que reemplaza 2 por 1 y 1 por 0.

4. Comenzamos creando la variable indicadora del maximo nivel educativo del hogar, que divide a los encuestados a partir de “codusu” y determina cual es el nivel educativo más

alto alcanzado en este hogar. Esta variable es indivadora del contexto socioeconómico de la familia y puede estar relacionada con la desocupación.

Luego creamos el indice de ayuda extra que a partir de las variables 'v5' (subsidio o ayuda social), 'v6' (mercaderías, ropa, alimentos del gobierno, iglesias, escuelas, etc.) y 'v12' (cuotas de alimentos o ayuda en dinero de personas) asigna un número que representa cuantos de estos tres tipos de ayuda externa recibe el encuestado. Esta variable se relaciona con la descoupación ya que indica en que medida la persona depende de ayudas externas por algún tipo de precariedad o ausencia laboral.

Por último, creamos la variable indicadora de la cantidad de menores de 16 año, ya que éste grupo étario, por lógicas razones, en la gran mayoría de los casos no tienen trabajo (ni lo buscan activamente). Con este objetivo dividimos a los encuestados por su “codusu” y contamos la cantidad de valora < 16 de la variable “ch06”.

5. Comenzamos analizando estadísticas descriptivas como la media, el desvío estandar y el máximo y mínimo de tres variables: edad (ch06), sabe leer (ch09) y recibe subsidios (v5), comparando los años 2004 y 2024. En cuanto a la edad, la media pasó de 33.67 años en 2004 a 38.47 años en 2024, indicando un envejecimiento de la población. La alfabetización (ch09) muestra una mejora significativa: el porcentaje de personas que saben leer aumentó del 89% en 2004 al 95% en 2024. Esto refleja avances en educación y puede relacionarse con la desocupación al aumentar las capacidades laborales de la población. Respecto a los subsidios gubernamentales (v5), solo el 6.3% de la población los recibía en 2004, cifra que creció al 18.1% en 2024. Este aumento podría estar vinculado a políticas sociales más amplias o a un mayor número de personas en condiciones de vulnerabilidad.

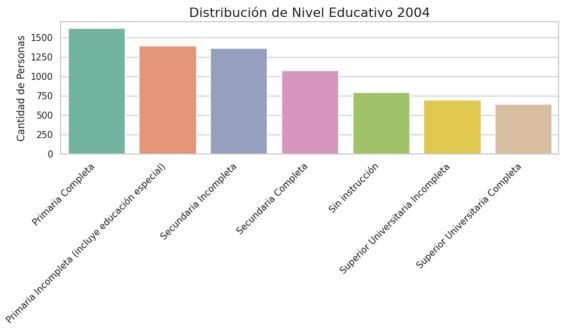
Estadísticas Descriptivas de Variables Relevantes para Predecir la Desocupación 2004

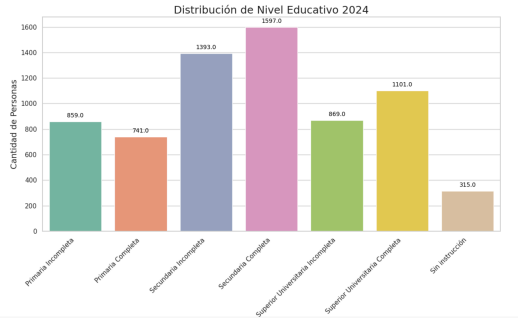
	ch06	ch09	v5
count	7609.000000	7609.000000	7609.000000
mean	33.672493	0.889736	0.063083
std	22.696941	0.313239	0.243129
min	1.000000	0.000000	0.000000
25%	15.000000	1.000000	0.000000
50%	30.000000	1.000000	0.000000
75%	50.000000	1.000000	0.000000
max	98.000000	1.000000	1.000000

Estadísticas Descriptivas de Variables Relevantes para Predecir la Desocupación 2024

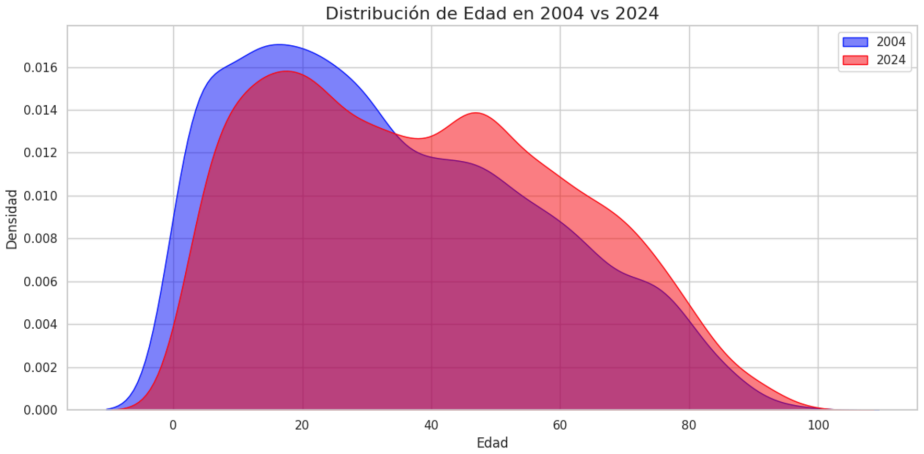
	ch06	ch09	v5
count	6903.000000	6903.000000	6903.000000
mean	38.472403	0.953788	0.181515
std	22.574647	0.209959	0.385472
min	2.000000	0.000000	0.000000
25%	19.000000	1.000000	0.000000
50%	37.000000	1.000000	0.000000
75%	56.000000	1.000000	0.000000
max	97.000000	1.000000	1.000000

Por otro lado, realizamos gráficos para observar el comportamiento de las variables. En cuanto a la distribución de nivel educativo se vieron mejoras considerables en el nivel general de la población. En 2004 la mayoría de la población únicamente había completado la primaria y el grupo más reducido eran los estudiantes universitarios. Esto merjoró drásticamente en 2024, donde la gran mayoría de las personas terminaron el secundario y la minoría de personas son quienes no tienen ningún tipo de instrucción.



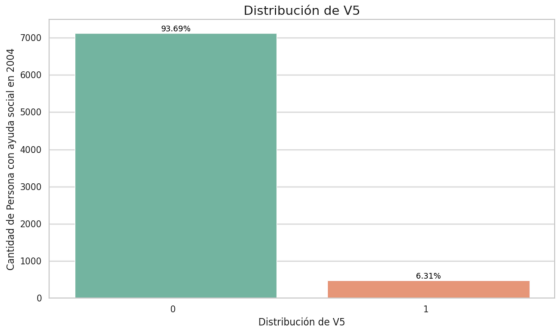


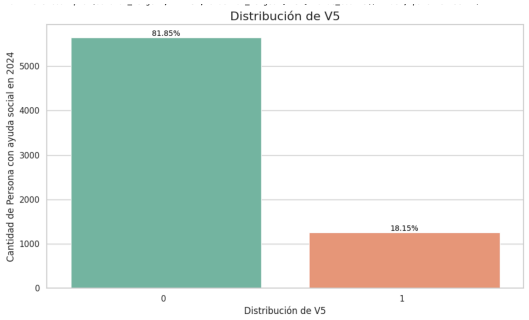
A continuación analizamos la distribución de la edad en ambos años. En 2004 la densidad alcanza su punto máximo entre los 20 y 30 años, sugiriendo que en 2004 la mayor concentración de personas se encontraba en este rango de edad. En 2024, el pico de la distribución parece desplazarse ligeramente hacia edades mayores (entre los 30 y 40 años), indicando un leve envejecimiento de la población respecto a 2004. Además, hay una mayor densidad en edades avanzadas (50+ años) en comparación con 2004. Por últimom, en 2024 la curva es más ancha, lo que indica una distribución más homogénea entre diferentes grupos etarios, mientras que en 2004 estaba más concentrada en personas jóvenes.



En la variable “V5” se observa un cambio significativo en la cantidad de gente que recibe subsidios, pasando del 6.31% al 18.15%. Este cambio se puede adjudicar tanto a estar

vinculado a políticas sociales más amplias o a un mayor número de personas en condiciones de vulnerabilidad.





PARTE 2: Clasificación y regularización

1. En primer lugar, creamos la variable “desocupados”, que posteriormente estableceremos como la variable dependiente en el conjunto de entrenamiento (vector y), mientras que el resto de las variables serán consideradas como las variables independientes (matriz X). Además, antes de realizar la división de la base transformamos las variables categoricas en dummies.

Una vez preparada la base, utilizamos el comando train_test_split, asignando el 70% de los datos al conjunto de entrenamiento y el 30% al conjunto de prueba, con una semilla de 101 para garantizar la reproducibilidad de los resultados (Nguyen, 2020). La separación de los datos en conjuntos de entrenamiento y prueba es crucial en el desarrollo de modelos de machine learning. El conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba sirve para evaluar su rendimiento en datos no vistos. Esto permite detectar problemas como el sobreajuste, en el cual un modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos (Bishop, 2006). Posteriormente, establecemos la variable "desocupado" como la variable dependiente en el conjunto de entrenamiento (vector y), mientras que el resto de las variables serán consideradas como las variables independientes (matriz X).

2. Se emplea la técnica de validación cruzada (CV) para evaluar diversos valores de λ y elegir el que minimice el error de validación. En este enfoque, los datos se dividen en k subconjuntos, utilizando uno para validación y el resto para entrenamiento. Para cada valor de λ, el modelo se ajusta k veces, entrenando con k-1 particiones y evaluando con la partición restante en cada iteración. Posteriormente, se calcula el error para cada partición y se obtiene el error promedio, conocido como CV. Este proceso se repite para distintos valores de λ, seleccionando el que minimiza el MSE de la validación cruzada. Es importante resaltar que

este procedimiento solo se aplica al conjunto de entrenamiento para evitar sesgar la estimación del error utilizando el conjunto de prueba. Si se usara el conjunto de prueba, se correría el riesgo de sobreajustar el modelo a esos datos, afectando la capacidad del modelo para generalizar a nuevos datos.

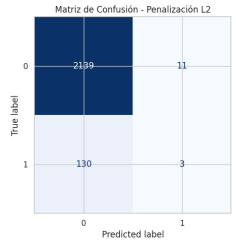
3. Las implicancias del tamaño del k se relacionan con el trade-off entre el sesgo y la varianza. Con un K pequeño las particiones van a tener más datos pero menos variabilidad entre la partición de entrenamiento y validación, elevando el sesgo. Por otro lado, un *k* muy grande aumenta la varianza. Al usar Leave-one-out CV, cuando se elige que $k = n$ (n muestras) cada iteración utiliza n-1 observaciones para entrenar el el método de aprendizaje. Esto que implica que el modelo se estima n veces, una por cada observación. Esto genera n estimaciones del error, dando como resultado n valores del MSE.

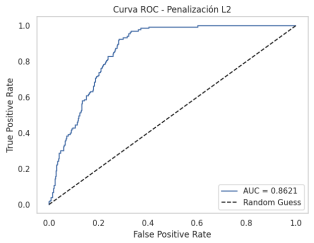
4. Para la regresión logística, implementamos la penalización L1 (como la de Lasso) y L2 (como la de Ridge) con lambda = 1 (como en el Tutorial 10), utilizando la opción penalty en el modelo. Primero estandarizamos y entrenamos el modelo con penalización L1, lo que permitió realizar una selección de características al forzar algunos coeficientes a cero, y luego con penalización L2, que ayudó a reducir la magnitud de los coeficientes sin llevarlos a cero. Después de entrenar los modelos con ambas penalizaciones, reportamos la matriz de confusión, la curva ROC, y los valores de AUC y Accuracy para cada año, evaluando cómo se desempeñó el modelo en función de estas métricas. Estos resultados nos permitieron comparar el rendimiento de los modelos con penalización L1 y L2, proporcionando una visión detallada de su capacidad de clasificación y generalización.

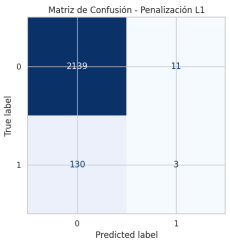
En el TP 3, para 2004, la regresión logística mostró una exactitud del 97.01% y un AUC de 0.989. Sin embargo, en 2024 el AUC cae a 0.64, lo que refleja una pérdida significativa en la capacidad del modelo para distinguir entre clases. Aunque mantiene pocos falsos positivos, el aumento de falsos negativos sugiere que el modelo es menos efectivo en la detección de casos positivos.

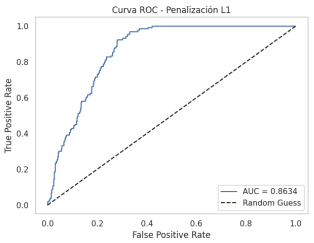
Para el año 2004 los resultados de regresión logística con penalizaciones L1 y L2 muestran un rendimiento similar entre sí. En el modelo con penalización L1, los valores son casi idénticos, con 3 verdaderos positivos, 2144 verdaderos negativos, 6 falsos positivos y 130 falsos negativos, y un AUC de 0.8639. Para el modelo con penalización L2, la matriz de confusión presenta 4 verdaderos positivos, 2143 verdaderos negativos, 7 falsos positivos y 129 falsos negativos, con un AUC de 0.8648, indicando buena capacidad discriminativa pero con un error en la clasificación de instancias positivas. En resumen, el modelo con

penalización L2 muestra un rendimiento ligeramente superior, pero las diferencias son mínimas.

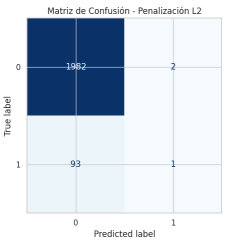


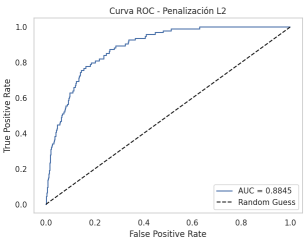


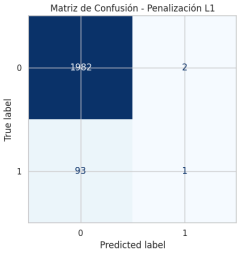


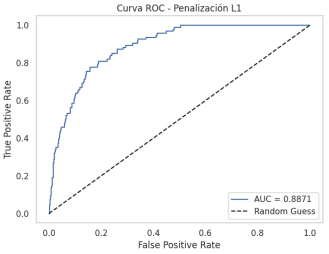


Para el año 2024 los resultados de regresión logística con penalizaciones L1 y L2 son identicos. La matriz de confusión presenta 2 verdaderos positivos, 1984 verdaderos negativos, 0 falsos positivos y 92 falsos negativos. Ambos muestran un AUC de 0.89, indicando buena capacidad discriminativa pero con un error en la clasificación de instancias positivas.



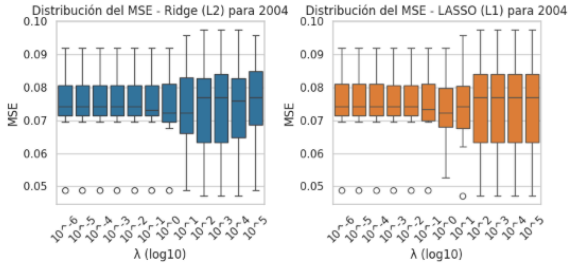


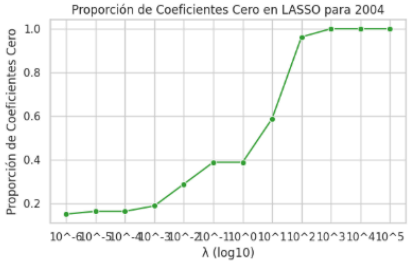


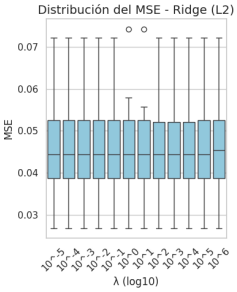


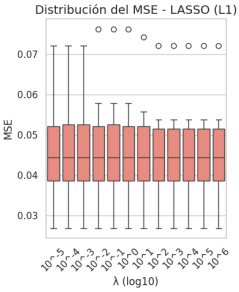
En conclusión, la performance general de la regresión logística con regularización parecería ser superior en 2024 en comparación con el TP3 en términos de AUC, sin embargo, el modelo tiene claras dificultades para clasificar correctamente instancias positivas, sobre todo en 2024, lo que evidencia grandes limitaciones. La elección entre L1 y L2 parece no ser determinante en este caso, ya que ambos ofrecen resultados muy similares.

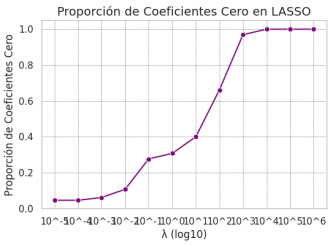
5. Se llevó a cabo un análisis en un rango de lambda = 10^n con $n \in [-6,6]$ usando 10-fold cross-validation para ambos modelos, LASSO y Ridge, en los años 2004 y 2024. El lambda óptimo se determinó seleccionando el que minimizó el error cuadrático medio. En el caso del año 2004, el valor seleccionado fue 10 para Ridge y 1 para LASSO. Para el año 2024, lambda óptimo también fue $\lambda = 100$ para Ridge y 10 para LASSO. Adicionalmente, se generaron boxplots que comparan las regularizaciones Ridge (L2) y Lasso (L1) en términos del error cuadrático medio y la proporción de coeficientes anulados en Lasso, para los periodos 2004 y 2024.











Para 2004 Ridge (L2) y LASSO (L1) muestran diferencias en su comportamiento respecto al MSE y la selección de variables. Mientras que Ridge mantiene el MSE estable sin importar el valor de lambda, LASSO presenta un aumento en el MSE a medida que lambda crece. LASSO también reduce rápidamente la cantidad de coeficientes no nulos al incrementar lambda, lo que lo hace útil para la selección de variables. En cambio, Ridge mantiene la estabilidad de los coeficientes sin eliminar variables.

En el caso de 2024, en Ridge el MSE se mantiene estable a lo largo de diferentes valores de lambda, lo que indica que este método no se ve significativamente afectado por la regularización. Por otro lado, en LASSO, el MSE presenta un leve aumento con valores más altos de lambda, evidenciando una pérdida de precisión cuando la penalización se intensifica. Además, el tercer gráfico destaca que en LASSO la proporción de coeficientes reducidos a cero crece rápidamente con el incremento de lambda, alcanzando el 100% en valores elevados. Esto confirma la capacidad de LASSO para realizar selección de variables eliminando las menos relevantes.

6 - En el caso del valor óptimo de λ para LASSO encontrado en el inciso anterior, se han descartado un total de 31 variables, lo que deja al modelo final con 49 variables con coeficientes distintos de cero, las cuales son consideradas relevantes para la predicción. Las otras variables han sido eliminadas porque LASSO las ha forzado a tener coeficientes iguales a cero, lo que implica que no aportan significativamente al modelo. Algunas de las variables descartadas no eran esperadas, como “Nivel_ed” (Nivel educativo) y “v12” (¿Cuotas de alimentos o ayuda en dinero de personas que no viven en el hogar?), mientras que otras,

como “cho4” (Sexo), “ch07” (Estado civil) y “ch08”, eran esperadas debido a que su influencia sobre la variable objetivo podría ser más limitada. En el análisis correspondiente al valor óptimo de λ para LASSO en el año 2024, se descartaron 26 variables, dejando un total de 39 características relevantes con coeficientes distintos de cero. Algunas de las variables eliminadas, como “ch04” (Sexo) y “ch07” (Estado civil), eran previsibles, pero otras, como “indice_ayuda_externa”, “menor_de_16” e “IV1” (Tipo de vivienda), nos sorprendieron al ser descartadas, ya que se pensaba que podrían ser útiles para predecir la desocupación.

En relación con las variables seleccionadas en el inciso 1 del punto 1, se conservaron algunas relevantes, como “v5” y “v6”, que indican si el hogar recibe subsidios o ayuda social (dinero, mercadería, ropa, alimentos). También se incluyó “v12”, que señala si el hogar recibió ayuda externa en forma de cuotas de alimentos o dinero. Además, se mantuvieron variables como “IX_TOT” (número de miembros del hogar) e “ITF” (ingreso total familiar), ya que son indicadores clave del nivel de ingresos del hogar y podrían estar directamente relacionados con la desocupación.

7 - Al analizar el error cuadrático medio (MSE) del parámetro alpha obtenido mediante validación cruzada para cada año, podemos determinar qué método de regularización (Ridge o Lasso) ha dado los mejores resultados. Un MSE más alto indica que el modelo tiene un peor desempeño, ya que refleja mayores diferencias entre las predicciones y los valores observados, lo que sugiere un mayor error en las predicciones.

Por un lado, al comparar los valores de Error Cuadrático Medio (ECM) en 2004, el modelo Ridge presenta un rendimiento ligeramente superior, ya que su ECM (0.0604) es más bajo que el del modelo Lasso (0.0613). Esto sugiere que Ridge tiene un mejor desempeño en este caso particular, al ser más preciso en la predicción de los datos. Por otro lado, el ECM para ambos modelos en 2024 es idéntico: ECM de Ridge 2024: 0.0457 y ECM de Lasso 2024: 0.0457. Esto indica que Ridge y Lasso presentan un rendimiento idéntico en este caso, ya que ambos modelos tienen un nivel de precisión similar en la predicción de los datos.

En conclusión, Ridge parece ser más robusto en 2004, mientras que en 2024 ambos modelos muestran una precisión similar, lo que sugiere que, en este caso, la selección de características de Lasso no aporta una ventaja significativa.

Bibliografía

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Diligent. (n.d.). *Método outliers() en HCL*. En *Centro de ayuda de la plataforma Diligent One*. Recuperado de https://help.highbond.com/helpdocs/highbond/es/Content/robots/scripting/hcl/hcl_outliers.htm

Nguyen, M. (2020). *A guide on data analysis*. Bookdown. https://bookdown.org/mike/data_analysis/