Universidad de
SanAndrés

Trabajo práctico 4

Clasificando de de desocupados en la EPH: métodos de regularización y CART

E337 - Big Data

María Noelia Romero
Tomás Pacheco

Rosario Namur, Marcos Ohanian & Juliana Rodas

Otoño 2025

Ejercicio 1

Al utilizar penalizaciones como LASSO y Ridge en nuestras estimaciones, es fundamental considerar el valor del parámetro λ, ya que este determina el grado en que se sesgarán los coeficientes del modelo. El objetivo es introducir cierto sesgo en los coeficientes con el fin de reducir drásticamente la varianza, logrando así minimizar el error de predicción total. Un mayor valor de λ implica un mayor sesgo en los coeficientes en comparación con las estimaciones obtenidas mediante MCO.

Una forma adecuada de elegir el parámetro λ es mediante *cross-validation*. Este procedimiento consiste en dividir el conjunto de entrenamiento en varios bloques o "folds". Luego, para cada valor posible de λ, se entrena el modelo utilizando $K-1$ folds y se evalúa el error de predicción sobre el fold restante. Este proceso se repite K veces, alternando los folds de entrenamiento y validación, y se promedia el error cuadrático medio (MSE) para cada valor de λ. El valor óptimo será aquel que minimice dicho error promedio. Finalmente, se utiliza el valor de λ obtenido por validación cruzada para determinar el grado de penalización que cada método aplicará a los coeficientes.

Cuando separamos nuestra base de datos en conjuntos de entrenamiento y testeo, el objetivo es entrenar el modelo sobre la primera muestra y luego evaluarlo intentando predecir la segunda. Este enfoque permite identificar posibles problemas de *overfitting*, es decir, detectar si el modelo se ajusta en exceso a los datos de entrenamiento y por tanto pierde capacidad de predicción fuera de la muestra. El conjunto de prueba debe reservarse exclusivamente para evaluar el rendimiento final del modelo ya ajustado, simulando su comportamiento ante datos nuevos y no observados previamente. Si se utilizara el conjunto de prueba para seleccionar el parámetro λ, se introduciría un sesgo optimista en la estimación del MSE, ya que el modelo habría sido ajustado indirectamente sobre estos datos. Como consecuencia, se comprometería la validez del análisis predictivo, ya que no se estaría utilizando un conjunto verdaderamente independiente para validar los resultados del modelo.

Ejercicio 2

La elección del número de pliegues k en un esquema de *cross-validation* tiene importantes implicancias tanto en términos del sesgo como de la varianza de la estimación del error de *cross validation*.

Un valor de k muy pequeño puede generar una alta varianza en la estimación del error mediante *cross-validation*. Esto se debe a que los datos asignados a cada pliegue influyen significativamente en el resultado: al haber pocos folds, se realizan pocas iteraciones del procedimiento y se cuenta con una cantidad limitada de estimaciones del error de testeo. Esta escasez de repeticiones reduce la estabilidad del estimador y aumenta su varianza. En el caso extremo, cuando $k = 1$, se realiza un único split entre entrenamiento y testeo, lo que impide capturar adecuadamente la variabilidad del error de predicción.

En el extremo opuesto, cuando $k = n$, se utiliza el esquema conocido como *Leave-One-Out Cross-Validation*. En este caso, el modelo se entrena n veces, cada vez dejando una única observación fuera para testeo. Este enfoque permite aprovechar al máximo los datos para el entrenamiento (conjuntos de entrenamiento de tamaño $n-1$), lo qu reduce el sesgo de la estimación del error de predicción de *cross-validation*. No obstante, presenta dos desventajas importantes: por un lado, cada grupo de testeo contiene solo una observación, lo que incrementa la varianza de la estimación; por otro lado, el procedimiento es computacionalmente costoso, especialmente para muestras grandes.

En resumen, elegir un valor intermedio de k (por ejemplo, $k = 5$ o $k = 10$) suele ofrecer un buen equilibrio

entre sesgo, varianza y costo computacional. Es importante no elegir un k ni demasiado pequeño ni demasiado grande para obtener estimaciones confiables del desempeño del modelo.

Ejercicio 3

Ahora podemos aplicar a regresión logística ambos métodos de regularización. Presentamos las matrices de confusión, medida de *acurracy*, Curva ROC (Figura 8) y AUC para Lasso (L1) y Ridge (L2).

Table 1: Penalización de Lasso (por año)

	2004	**2024**
Matriz de confusión	$\begin{bmatrix} 2150 & 1 \\ 140 & 1 \end{bmatrix}$	$\begin{bmatrix} 2002 & 0 \\ 101 & 0 \end{bmatrix}$
Accuracy	0.9385	0.9520
AUC	0.8054	0.7473

Table 2: Penalización de Ridge (por año)

	2004	**2024**
Matriz de confusión	$\begin{bmatrix} 2150 & 1 \\ 140 & 1 \end{bmatrix}$	$\begin{bmatrix} 2002 & 0 \\ 101 & 0 \end{bmatrix}$
Accuracy	0.9385	0.9520
AUC	0.8057	0.7435

Al comparar las métricas solicitadas para Lasso y Ridge en las Tablas 1 y 2, podemos notar que no hay una diferencia muy notable en las medidas de *accuracy* y AUC; incluso la matriz de confusión resulta la misma. Ahora, analizamos los resultados obtenidos en el Trabajo Práctico 3, en la regresión logística sin aplicar métodos de regularización:

Table 3: Regresión Logística (sin regularizar, por año)

	2004	**2024**
Matriz de confusión	$\begin{bmatrix} 2148 & 3 \\ 141 & 0 \end{bmatrix}$	$\begin{bmatrix} 2002 & 0 \\ 101 & 0 \end{bmatrix}$
Accuracy	0.9372	0.9520
AUC	0.6912	0.5803

De este modo, notamos una mejora en la predicción tanto con Lasso como con Ridge. En específico, vemos que para 2004 caen la cantidad de falsos positivos y falsos negativos. Por otro lado, hay una ligera mejora en el coeficiente de *accuracy* y una mejora más significativa en el área bajo la curva que, para 2004 aumenta de $0,6912$ a $0,8054$ (Lasso) y $0,8057$ (Ridge); mientras que para 2024 aumenta de $0,5803$ a $0,7473$ (Lasso) y $0,7435$ (Ridge).

Así, concluimos que al sesgar los coeficientes podemos reducir la varianza de nuestras predicciones, lo que nos lleva a una mejor performance del modelo de regresión logística.

Ejercicio 4

Ahora, realizamos un barrido de una lista de λ y utilizando cross-validation (con 10 folds) elegimos la penalización óptima para Ridge y Lasso. De esta manera, obtenemos que los λ que minimizan el MSE promedio son:

- Ridge 2004: 100
- Ridge 2024: 100
- Lasso 2004: 100

- Lasso 2024: 10

En esta línea, analizamos el promedio de MSE (en la Figura 9 del anexo) obtenido en la validación cruzada para cada λ propuesto.

En todos los gráficos vemos que el MSE promedio disminuye abruptamente cerca de $\lambda = 10$ y luego se estabiliza. A simple vista, podemos comprobar en Lasso 2004 y 2024 el mínimo alcanzado podemos ver que coincide con el λ antes mencionado.

Respecto a los gráficos de Ridge 2004 y 2024, parece en el gráfico que la penalización que minimiza es 10 cuando antes obtuvimos 100. Esta aparente discrepancia en se explica por la mínima diferencia en los valores del MSE entre $\lambda = 10$ y $\lambda = 100$, es tan pequeña que resulta imperceptible en los gráficos.

Por otro lado, una característica de Lasso es que este tipo de penalización puede hacer que algunos coeficientes sean exactamente cero. Así, Lasso también funciona como un método de selección de variables que, a mayor λ (es decir, más penalización) mayor cantidad de variables pasan a ser irrelevantes. Así, podemos graficar las variables descartadas para cada año con este método:

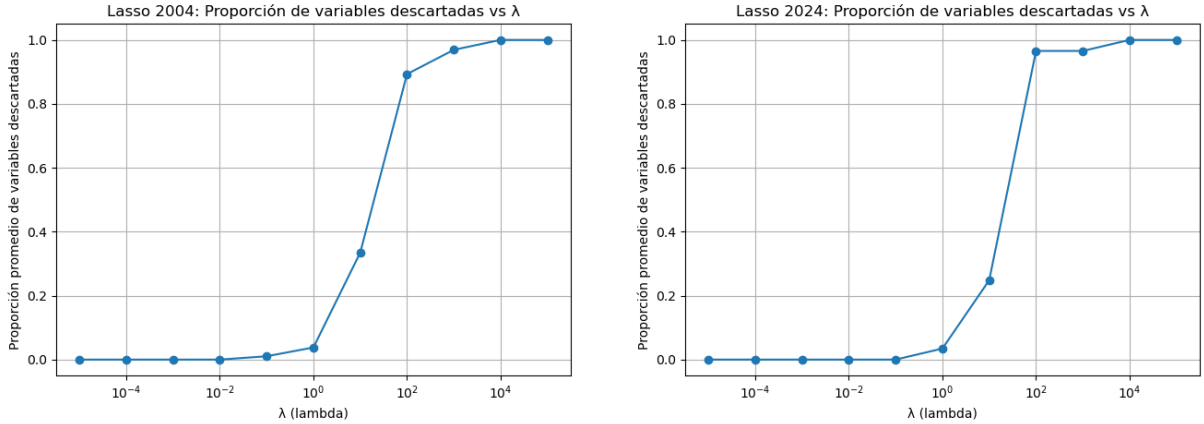


Figure 1: Proporción de variables descartadas por Lasso en 2004 y 2024

De esta manera, vemos que efectivamente mientras mayor sea λ menos variables va a utilizar el modelo (en otras palabras, tiene una mayor proporción de variables descartadas). En este caso, vemos que la mayoría de las variables las descarta entre $\lambda = 10^0$ y $\lambda = 10^2$.

Ejercicio 5

En primer lugar, consideraremos los resultados para el año 2004. Las variables descartadas son: edad, edad al cuadrado, la proporción del ingreso total del hogar que proviene de fuentes no laborales, la proporción de educación en la vida, mujeres con hijos, hijos por edad, la relación con el jefe de hogar (es decir, todas las dummies creadas para la variable categórica $ch03$), el sexo, el estado civil, las capacidades lingüísticas del individuo, la dummie que indica que el individuo nunca asistió a algún establecimiento educativo ($ch10_3$), movilización laboral y responsable único del hogar. Las variables restantes (aquellas que no fueron sesgadas a cero) son: la constante, el monto de ingreso per cápita familia, los años de educación y la dummie que indica que el individuo asistió en el pasado a algún establecimiento educativo ($ch03_2$).

En segundo lugar, consideramos los resultados para el año 2024. Las variables descartadas son: la edad, la dummie que indica "otros familiares" en cuanto a la relación con el jefe de familia ($ch03_9$), el sexo, si el individuo es soltero o no, si el individuo nunca asistió a un establecimiento educativo y responsable único del hogar. Las variables restantes (aquellas que no fueron sesgadas a cero) son: la constante, el monto de ingreso pér cápita familia, la edad al cuadrado, la proporción del ingreso total del hogar que proviene de fuentes no laborales, los años de educación, la proporción de educación en la vida, mujeres con hijos, hijos por edad, todas las relaciones de parentesco con el jefe de familia (exceptuando a "otros familiares"), el estado civil del individuo, las capacidades lingüísticas del individuo, si el individuo asistió en el pasado a algún establecimiento educativo y movilización laboral.

Los resultados obtenidos son congruentes con los valores óptimos de λ obtenidos para cada año. Debido a que los valores de λ que minimizan el MSE promedio con el método de LASSO son 100 para 2004 y 10 para 2024, tiene sentido que haya una mayor cantidad de variables descartadas para el primer año. Podemos observar que 25 variables fueron descartadas para 2004, mientras que, tan solo 6 variables para 2024. A pesar de lo recién mencionado, podemos resaltar patrones interesantes. En primer lugar, el sexo fue descartados para ambos años. Esto llama la atención pues es posible suponer que el sexo sería una variable importante a tener en cuanta al momento de estimar si una persona es ocupada o desocupada. En segundo lugar, el monto de ingreso pér cápita familia, los años de educación y la dummie que indica que el individuo asistió en el pasado a algún establecimiento educativo son variables que no fueron descartas en ninguno de los dos años. En un comienzo, suponíamos que estos dos primeros regresores iban a ser relevantes en ambos casos, sin embargo, encontramos llamativo el hecho de que la condición de los individuos de haber asistido en el pasado a un establecimiento educativo haya sobrevivido en el primer caso a pesar de la gran penalización impuesta.

Ejercicio 6

Luego de utilizar la especificación con el valor óptimo de λ para cada uno de los métodos de regularización (LASSO y Ridge), se evaluó su desempeño en términos del error cuadrático medio (MSE) sobre el conjunto de testeo, tanto para el año 2004 como para el año 2024. A continuación, se presenta una tabla con los resultados obtenidos:

Método	**MSE testeo 2004**	**MSE testeo 2024**
LASSO	0.0574	0.0442
Ridge	0.0548	0.0450

Table 4: Error cuadrático medio (MSE) en testeo para LASSO y Ridge

Como se observa, ambos métodos presentan desempeños similares en términos del MSE. Para el año 2004, Ridge logra un menor MSE en comparación con LASSO (0.0548 frente a 0.0574), por lo que se considera el método preferido para ese año. En cambio, para el año 2024, LASSO presenta un MSE levemente inferior al de Ridge (0.0442 frente a 0.0450), por lo que se destaca como el método más eficiente para ese período. Esta diferencia sugiere que la capacidad predictiva relativa de cada método puede variar según el contexto temporal y las características de los datos.

Ejercicio 7

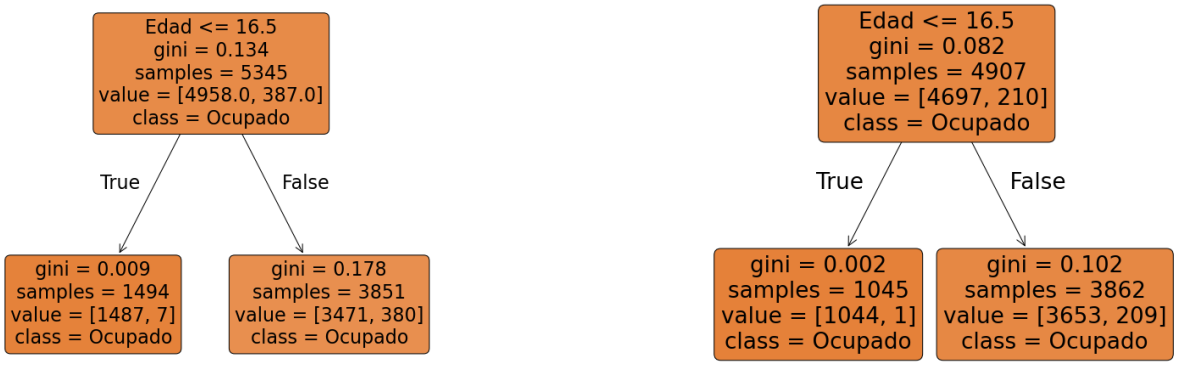
Al tratarse de un modelo de clasificación, no resulta adecuado estimar un árbol extenso para luego podarlo en función de la medida de costo de complejidad, como suele hacerse en regresión. En su lugar, seleccionamos el hiperparámetro correspondiente a la profundidad máxima del árbol (`max_depth`) utilizando validación cruzada con 10 particiones (*10-fold Cross-Validation*).

Para cada valor de `max_depth`, se entrena un modelo de árbol de decisión y se calcula el error de predicción en cada uno de los folds. Luego, se obtiene el error promedio y se selecciona la profundidad que minimiza este valor. A continuación, se presenta la *accuracy* promedio alcanzada para profundidades de 1 a 10, correspondiente a los años 2004 y 2024:

Max Depth	Accuracy 2004	Accuracy 2024
1	0.93	0.96
2	0.93	0.96
3	0.93	0.96
4	0.93	0.96
5	0.92	0.96
6	0.92	0.95
7	0.92	0.95
8	0.91	0.95
9	0.91	0.95
10	0.91	0.94

Table 5: Precisión promedio (accuracy) por profundidad máxima para los años 2004 y 2024

Si bien los resultados redondeados no permiten observar grandes diferencias, los valores exactos arrojados por el modelo en Python indican que la profundidad óptima es 1, con una *accuracy* de 0.9276 para el año 2004 y 0.9572 para el año 2024. Por lo tanto, los árboles finales resultan particularmente simples, como se muestra a continuación:



(a) Año 2004 (b) Año 2024

Figure 2: Árboles de decisión para los años 2004 y 2024

A partir de estos gráficos pueden extraerse varias conclusiones relevantes. En primer lugar, en ambos años el único predictor seleccionado es la edad (`ch06`), con un umbral de corte en 16.5 años. Esto sugiere

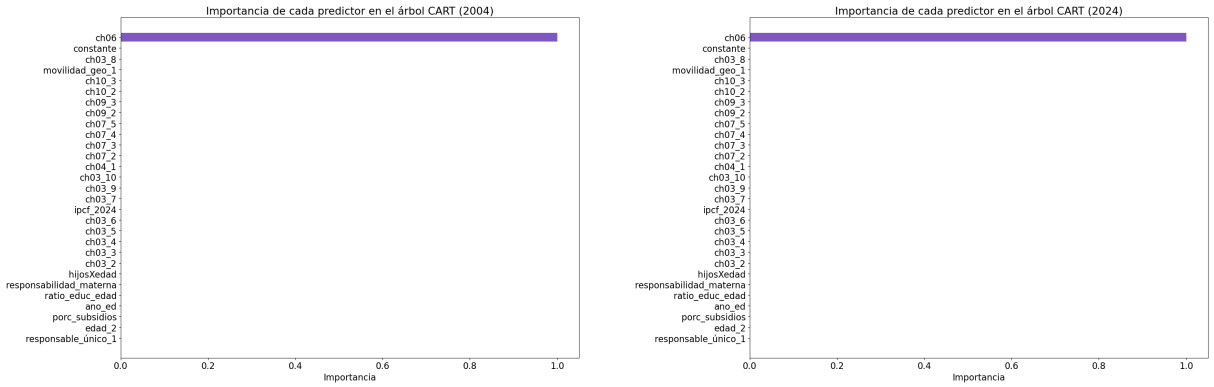
que la edad es, entre las variables disponibles, la más informativa para predecir el estado de ocupación. De hecho, el modelo clasifica como ocupadas tanto a personas mayores como menores de ese umbral, lo que puede deberse, en parte, al fuerte desbalance en la variable dependiente: los casos de desocupación son minoritarios en ambas muestras.

En segundo lugar, se observa que el índice de Gini en los nodos terminales es muy cercano a cero, lo cual es deseable. Esta métrica indica el grado de impureza de un nodo: valores bajos reflejan que las observaciones dentro del nodo son homogéneas en cuanto a su clase. Esto confirma que, pese a la simplicidad del árbol, las divisiones generadas son efectivas para separar los grupos.

Finalmente, cabe destacar que el hecho de que un árbol tan poco profundo alcance niveles altos de precisión sugiere que la clasificación es relativamente sencilla con las variables disponibles, pero también alerta sobre posibles limitaciones del modelo: al depender de una única variable, podría estar dejando sin captar otros aspectos relevantes a la hora de predecir el estado de ocupación.

Ejercicio 8

A continuación, mostramos en la Figura 3 la importancia que le da CART a cada uno de los predictores.



(a) Año 2004 (b) Año 2024

Figure 3: Importancia asignada por CART para cada uno de los predictores en 2004 y 2024

Tal como ambas figuras muestran, podemos observar que CART asignó toda la importancia únicamente a la edad de las personas. Tiene sentido que solo una variable tenga atribuida toda la importancia pues construimos ambos árboles con una profundidad de uno. Sin embargo, lo llamativo es que esta variable fue descartada para ambos años por LASSO con el método de LASSO . Tal como hemos discutido a lo largo de la asignatura, entendemos que la edad debe ser uno de los predictores más influyentes (junto con el ingreso y la educación) del estado de ocupación. Entonces, o que más llama nuestra atención es que este regresor no haya sido considerado por LASSO.

Ejercicio 9

Finalmente, calculamos la matriz de confusión, accuracy, Curva ROC (Figura 10) y AUC para cada árbol, así podemos comparar entre años y con los resultados obtenidos con los distintos métodos de regularización.

Table 6: Árbol de decisión (por año)

	2004	**2024**
Matriz de confusión	$\begin{bmatrix} 2151 & 0 \\ 141 & 0 \end{bmatrix}$	$\begin{bmatrix} 2002 & 0 \\ 101 & 0 \end{bmatrix}$
Accuracy	0.9385	0.9520
AUC	0.6412	0.6042

En particular, en la Tabla 6 podemos ver que en ambos casos tiene una alta tasa de falsos negativos (en efecto, el modelo no tiene una buena performance prediciendo a los desocupados pues es la categoría minoritaria). Luego, si pensamos en accuracy y AUC, CART tiene un mejor desempeño en 2004 que en 2024, de acuerdo a estas medidas.

Por otro lado, si comparamos estos resultados con los de Lasso y Ridge (Tablas 1 y 2, respectivamente) vemos que estos últimos tienen una performance ligeramente mejor a la hora de predecir a los desocupados, con 140 falsos negativos (a diferencia de 141 FN) en la matriz de confusión, en cada caso para 2004. Luego, en términos de accuracy los tres modelos tienen ,aproximadamente, la misma proporción de predicciones correctas. Sin embargo, si analizamos AUC esta medida es considerablemente mayor en los modelos regularizados que en el arbol (alrededor de 0,6 para CART -incluso menor que la regresion logistica sin regularizar- y alrededor de 0,8 para Lasso y Ridge). Esto significa que los modelos con métodos de regularización tienen una mayor capacidad de clasificar correctamente para diferentes umbrales.

En términos del trade-off entre comunicación y performance predictiva, el árbol de decisión (CART) ofrece una ventaja clara en cuanto a interpretabilidad ya que transmite fácilmente cómo se llega a una predicción. Sin embargo, como analizamos en líneas anteriores, esta simplicidad viene con un costo en la performance que se refleja en el bajo AUC. Esto indica que CART discrimina peor entre clases, es decir, su capacidad para generar predicciones probabilísticas informativas es más limitada. Por el contrario, Lasso y Ridge, aunque menos intuitivos para comunicar, logran una mejor performance predictiva.

Por último, también podemos comparar los errores de predicción entre los modelos y los años:

Modelo	**MSE 2004**	**MSE 2024**
Árbol	0.0615	0.0480
Lasso*	0.0574	0.0442
Ridge**	0.0548	0.0450

* $\lambda = 100$ en 2004 y $\lambda = 10$ en 2024.
** $\lambda = 100$ en ambos años

Table 7: Comparación de MSE de testeo por modelo y año

Con todas estas variables a considerar, podemos pensar en si hay una ventaja en aplicar un un método no lineal e interpretable como CART frente a los modelos lineales con penalización con menor interpretabilidad. En este caso, dado que el MSE es muy similar entre los modelos, el uso del árbol podría resultar conveniente por su mayor interpretabilidad. Sin embargo, al considerar otros umbrales de decisión, los modelos con regularización (Lasso y Ridge) muestran una performance predictiva mejor, reflejado en un mayor AUC. Bajo esta perspectiva, la ganancia en capacidad predictiva supera la ventaja interpretativa del árbol, haciendo preferible el uso de métodos de regularización.

Anexo

Ejercicio 3

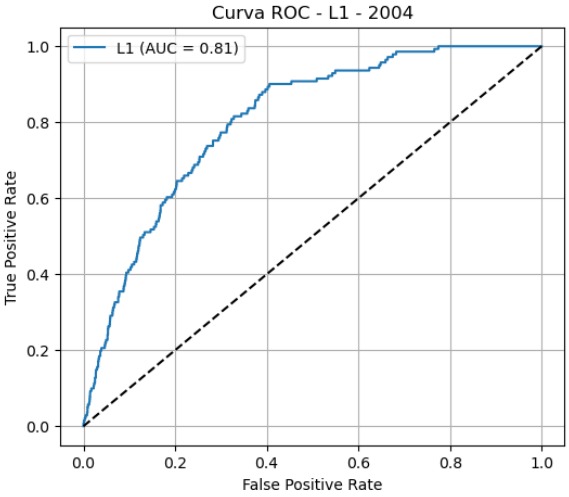


Figure 4: Penalización por Lasso (2004)

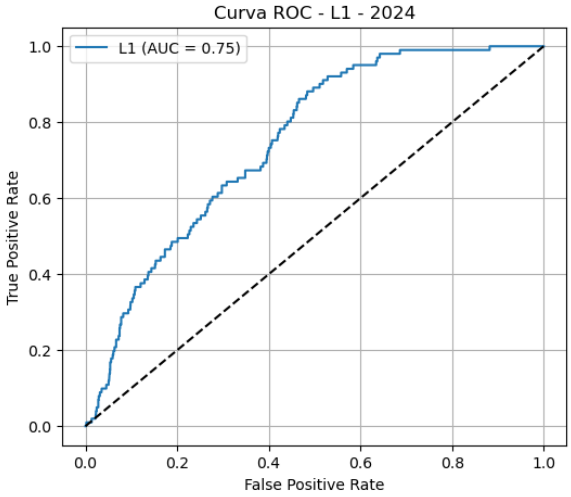


Figure 5: Penalización por Lasso (2024)

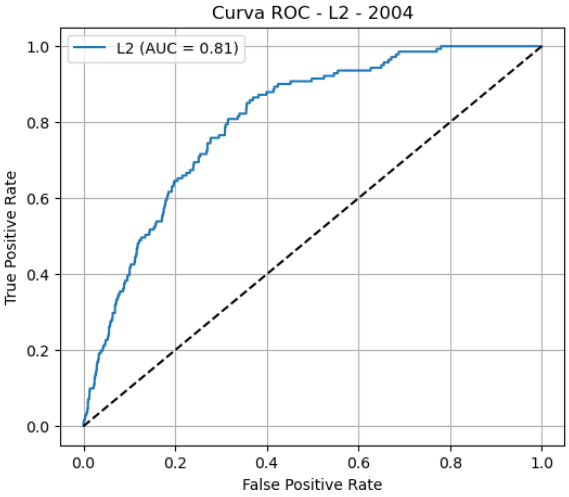


Figure 6: Penalización por Ridge (2004)

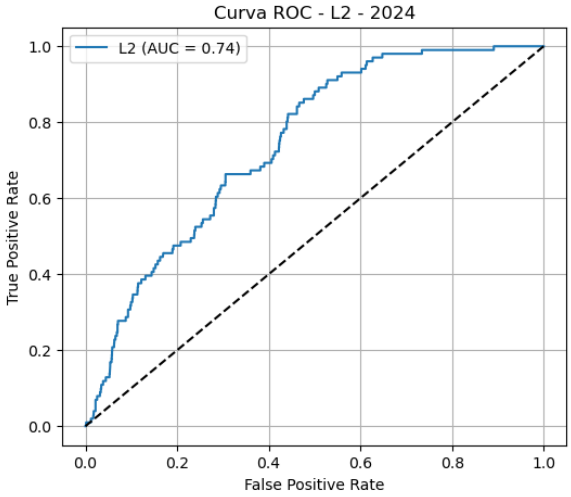


Figure 7: Penalización por Ridge (2024)

Figure 8: Curva ROC por año y método de regularización

Ejercicio 4

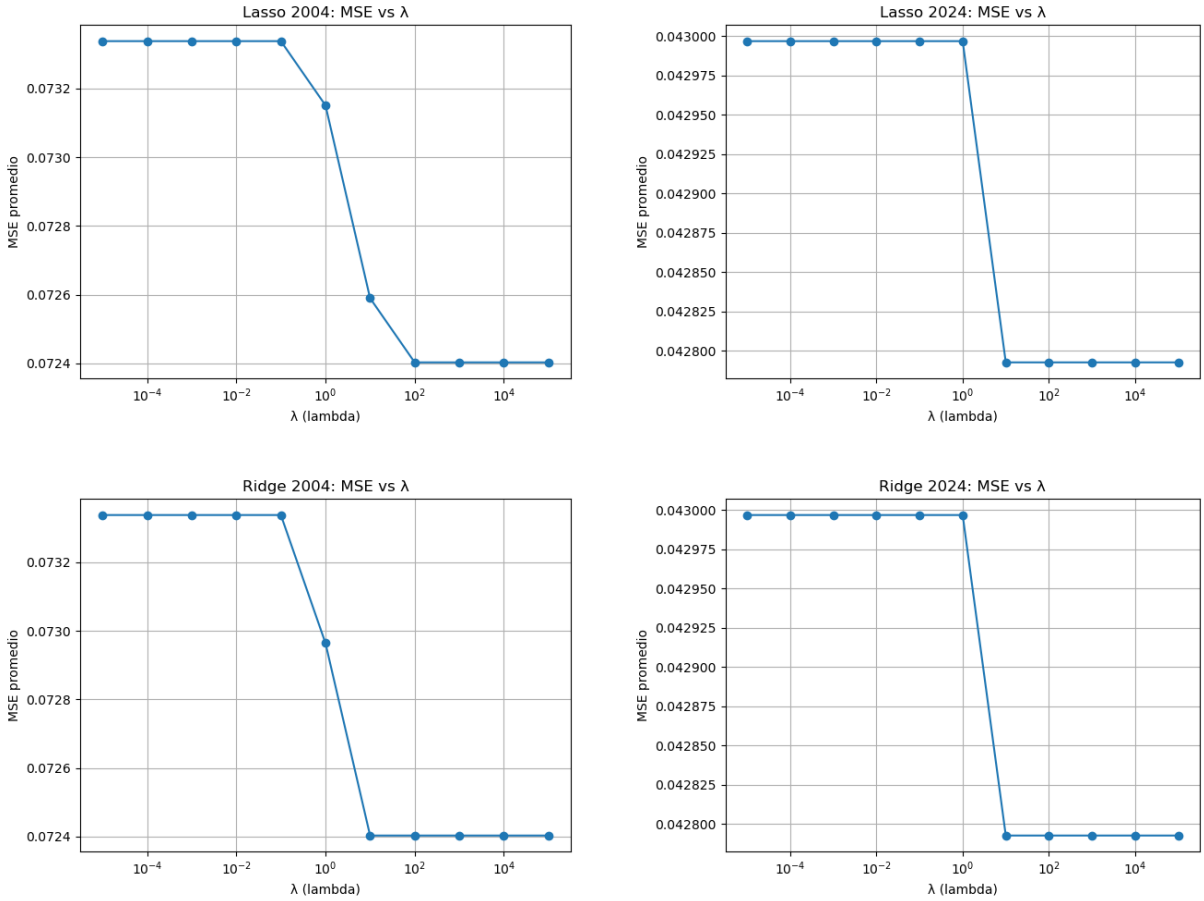


Figure 9: MSE de Lasso y Ridge para 2004 y 2024

Ejercicio 10

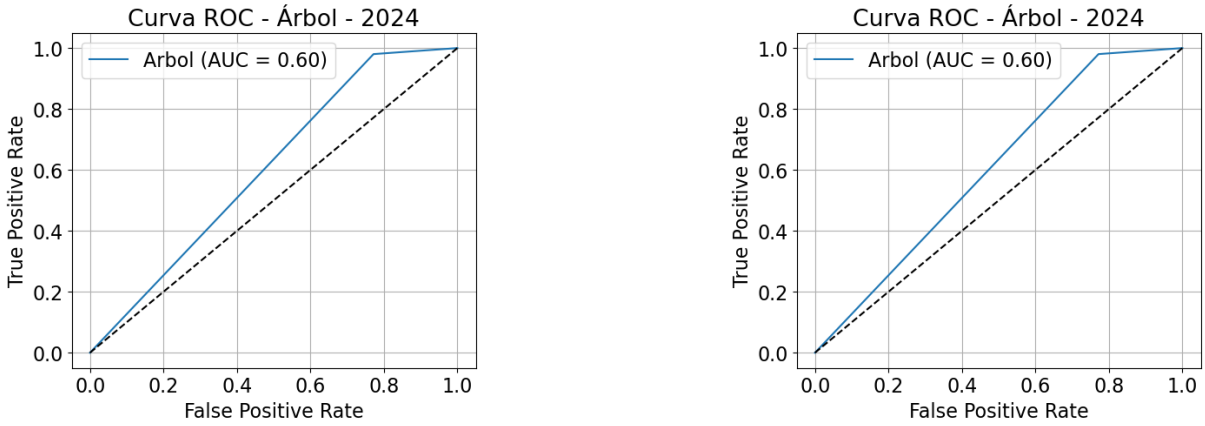


Figure 10: Curva ROC para árbol de decisión (2004 y 2024)