Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Maestría en Economía Aplicada

Materia:
Taller de Programación

Docente:
Noelia Romero

Informe – Trabajo Práctico # 4

Métodos de regularización y CART

Grupo # 5:

Cammisi, Andrés
Porco, Matías
Pineda, David

Link GitHub: https://github.com/DavidPT0902/Grupo_5_UBA_2025.git

Noviembre, 2025

1. Introducción

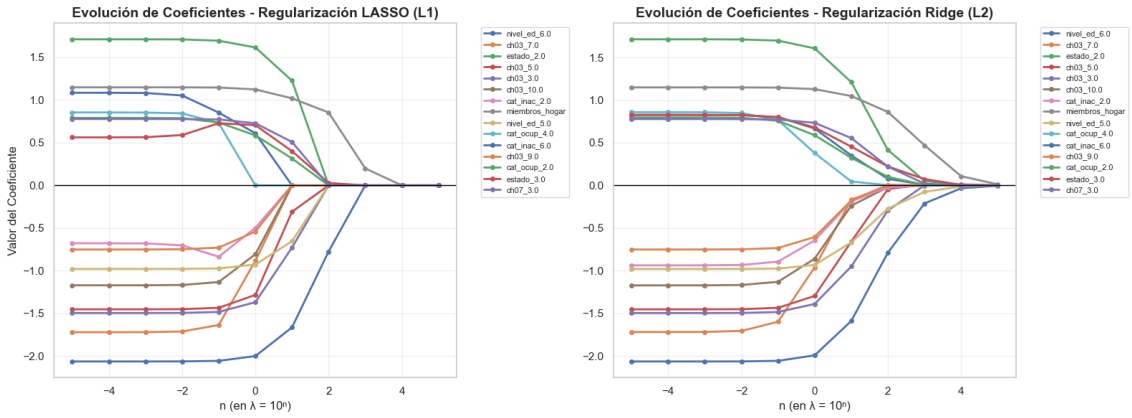
Predecir con precisión la pobreza es fundamental para diseñar políticas sociales efectivas. Este informe presenta los resultados del TP4, evaluando diferencias técnicas para predecir este fenómeno utilizando datos de la Encuesta Permanente de Hogares (EPH) 2025.

El documento se organizar en: 1) modelos de regresión logística con regularización Ridge y LASSO para controlar sobreajuste, 2) árbol de decisión podado para capturar relaciones no lineales, y3) comparación de modelos según su desempeño predictivo e interpretabilidad. El análisis utiliza 9 variables sociodemográficas y laborales (36 features) sobre 13,803 hogares, con división 70%-30% para entrenamiento y prueba.

2. Modelo de Regresión Logística con Regularización: Ridge y LASSO

2.1 Visualización

En esta sección analizamos la evolución de los coeficientes de la regresión logística penalizada. Para ello estimamos modelos con penalización L1 (LASSO) y L2 (Ridge) utilizando una grilla de valores del parámetro de penalidad de la forma $\lambda = 10^n$, con $n \in \{-5, \ldots, 5\}$. Dado que la implementación de LogisticRegression en sklearn utiliza $C = 1/\lambda$, valores pequeños de λ implican una penalización débil (C grande), mientras que valores elevados de λ corresponden a una penalización fuerte (C pequeño).



Para valores muy pequeños de λ (10^{-5} a 10^{-2}), los coeficientes prácticamente no difieren de los estimados por el modelo sin regularización. Esto ocurre porque la penalización es casi nula y el modelo reproduce el comportamiento del logit estándar. A partir de valores intermedios de penalización ($\lambda \approx 10^0 = 1$), LASSO comienza a ejercer un efecto apreciable. Algunos coeficientes se reducen rápidamente y varios de ellos se igualan exactamente a cero. Este comportamiento refleja la propiedad distintiva de la penalización L1, que no solo contrae los coeficientes, sino que también realiza una selección automática de variables. Finalmente, para valores grandes de

penalidad ($\lambda \geq 10^2$), todos los coeficientes colapsan a cero, lo que indica que la penalización domina completamente la función de pérdida y el modelo se vuelve no informativo.

Como puede observarse, el panel correspondiente a Ridge exhibe un comportamiento muy diferente. Al igual que en LASSO, cuando λ es muy pequeño los coeficientes prácticamente no cambian. Sin embargo, al aumentar la penalidad, la contracción de los coeficientes es suave y continua, sin quiebres abruptos y, además, sin que ningún coeficiente se vuelva exactamente cero. En efecto, para λ muy elevados (10^3 a 10^5), todos los coeficientes se acercan asintóticamente a cero, pero ninguno es eliminado. Esto confirma la principal propiedad de Ridge, que regulariza el modelo contrayendo los coeficientes, pero nunca conlleva una selección automática de variables.

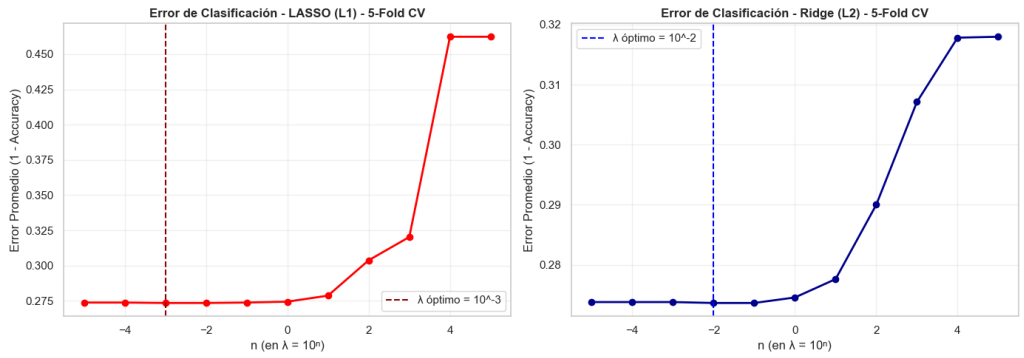
2.2Penalidad óptima por Cross-validation y visualización

Con el objetivo de elegir el nivel óptimo de regularización para los modelos logísticos penalizados, en esta sección aplicamos validación cruzada (CV) de 5 folds sobre la misma grilla utilizada previamente ($\lambda = 10^n$, con $n \in \{-5, \ldots, 5\}$). Para cada valor de λ, estimamos el error promedio de clasificación (1 – accuracy) y seleccionamos el parámetro que minimiza dicho error.

Tabla 1. Parámetro óptimo

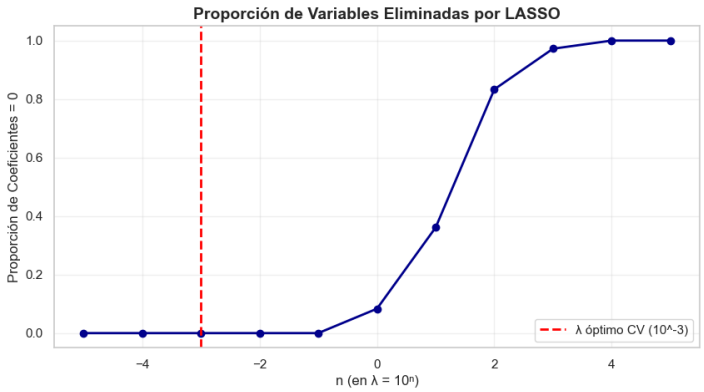
Método	λ Óptimo	C Óptimo	Error CV
LASSO (L1)	0.001 (10^{-3})	1000.0	27.35%
Ridge (L2)	0.001 (10^{-2})	100.0	27.36%

Para λ muy pequeños (10^{-5} a 10^{-2}), el error de clasificación para LASSO permanece prácticamente constante y cercano al mínimo. A partir de $\lambda \geq 10^0$, el error aumenta de manera marcada, coincidiendo con el inicio del proceso de eliminación de variables. Para λ grandes ($\geq 10^2$), el error se dispara, reflejando que el modelo queda excesivamente penalizado y pierde capacidad predictiva.



El mínimo error de clasificación para LASSO se obtiene en $\lambda = 10^{-3}$ (C=1000) con un error promedio de 0,2735. Observamos que, en este valor óptimo de penalidad, LASSO no elimina ninguna variable (0 de 36). Esto

indica que, en este caso, la regularización L1 mejora marginalmente el ajuste sin inducir un modelo más esparso.



El comportamiento del error de clasificación para Ridge es similar en forma, aunque presenta una trayectoria más suave. Para λ muy pequeños (10^{-5} a 10^{-3}), el error permanece estable. El mínimo se alcanza en $\lambda = 10^{-2}$. A partir de $\lambda \geq 10^{0}$, el error aumenta gradualmente, reflejando un exceso de contracción de los coeficientes. El mínimo error para Ridge se obtiene en $\lambda = 10^{-2}$ (C=100) con un error promedio de 0,2736. El modelo nunca elimina variables, consistente -como vimos- con las propiedades de la penalización L2.

Los resultados de CV permiten concluir que los dos modelos logran prácticamente el mismo error mínimo (LASSO = 0,2735, Ridge = 0,2736). En efecto, la diferencia es muy pequeña, por lo que la elección entre L1 y L2 no afecta de manera sustantiva la performance predictiva sobre este conjunto de datos.

2.3Estimación con λ^{cv} y comparación de coeficientes

En esta sección estimamos tres modelos logísticos utilizando el conjunto de entrenamiento:

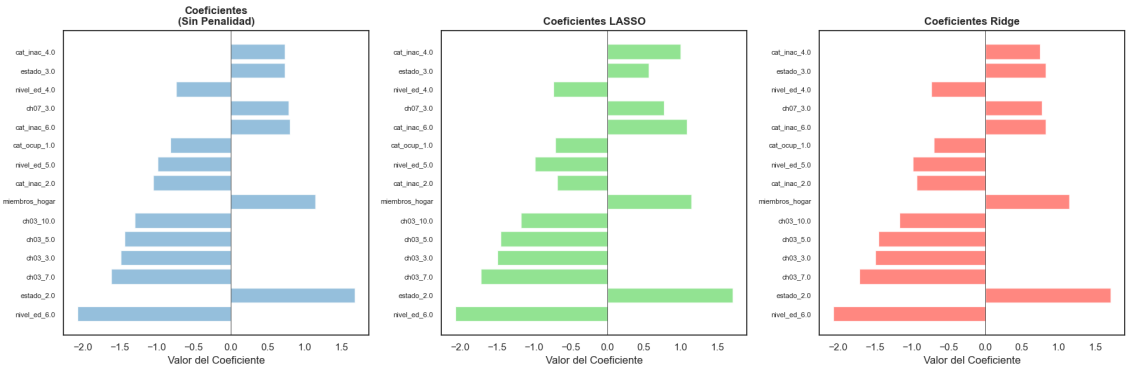
- Logit sin penalización.
- Logit con penalización L1 (LASSO) utilizando la penalidad óptima seleccionada por CV ($\lambda = 10^{-3}$).
- Logit con penalización L2 (Ridge) utilizando su penalidad óptima ($\lambda = 10^{-2}$).

Observando los resultados, comparamos los coeficientes estimados, la magnitud de los cambios introducidos por cada tipo de regularización y la eventual eliminación de variables.

Para empezar, observamos que ninguna variable fue eliminada por LASSO en la penalidad seleccionada por CV. Esto significa que, aunque LASSO tiene capacidad para fijar coeficientes iguales a cero, la validación cruzada determinó que la penalidad que minimiza el error de clasificación es muy débil como para inducirnos a la elección de un modelo más esparso. Por

lo tanto, los modelos estimados con L1 y L2 utilizan exactamente las mismas 36 variables que el logit sin penalización.

Los siguientes gráficos permiten comparar visualmente los tres modelos.



Como puede observarse, LASSO (L1) genera cambios algo mayores que Ridge, pero moderados. El cambio promedio absoluto con respecto al modelo sin penalización es igual a 0,091. No observamos diferencias en los signos ni en la importancia relativa de las variables. Algunas variables muestran contracciones relativamente fuertes (p. ej., cat_inac_2.0, estado_3.0, ch03_10.0), aunque sin llegar a cero.

Por su parte, Ridge (L2) aplica un grado de contracción más suave. En este caso, el cambio promedio absoluto es de 0,0367. Esto es consistente con el efecto típico de la penalización L2, que reduce la varianza del modelo sin inducir sparsity. En todos los casos, los coeficientes Ridge mantienen su signo y orden aproximado respecto al logit sin penalizar.

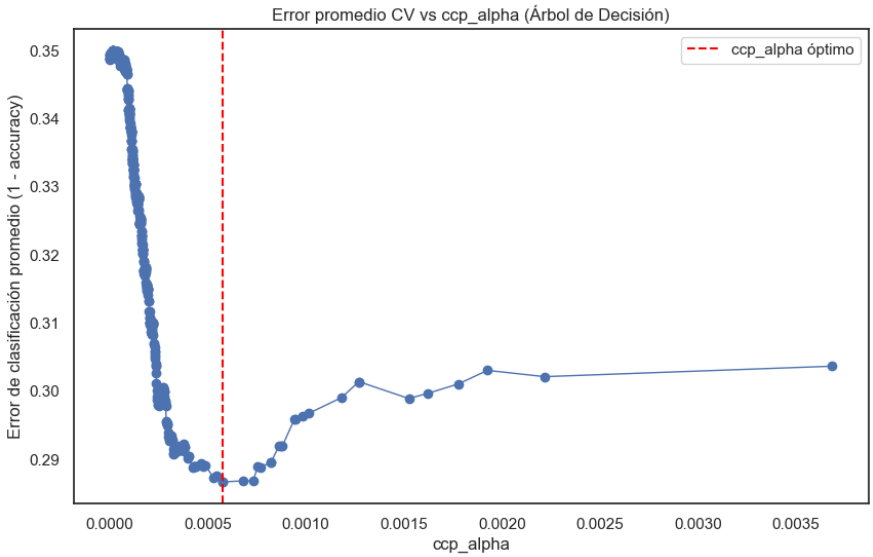
Los coeficientes más relevantes del modelo sin penalizar (como nivel_ed_6.0, estado_2.0, ch03_7.0, ch03_3.0, etc.) se mantienen como los predictores de mayor magnitud tanto en LASSO como en Ridge. Por lo tanto, las dos regularizaciones preservan la estructura central del modelo logit sin penalizar. En otras palabras, los predictores dominantes siguen siendo los mismos, y la regularización solo introduce pequeños ajustes en las magnitudes. En síntesis, la regularización mejora ligeramente la robustez y estabilidad del modelo, pero no altera su estructura fundamental ni selecciona variables adicionales, dado que los niveles de penalidad que optimizan la capacidad predictiva son relativamente bajos.

3. Árboles de Decisión

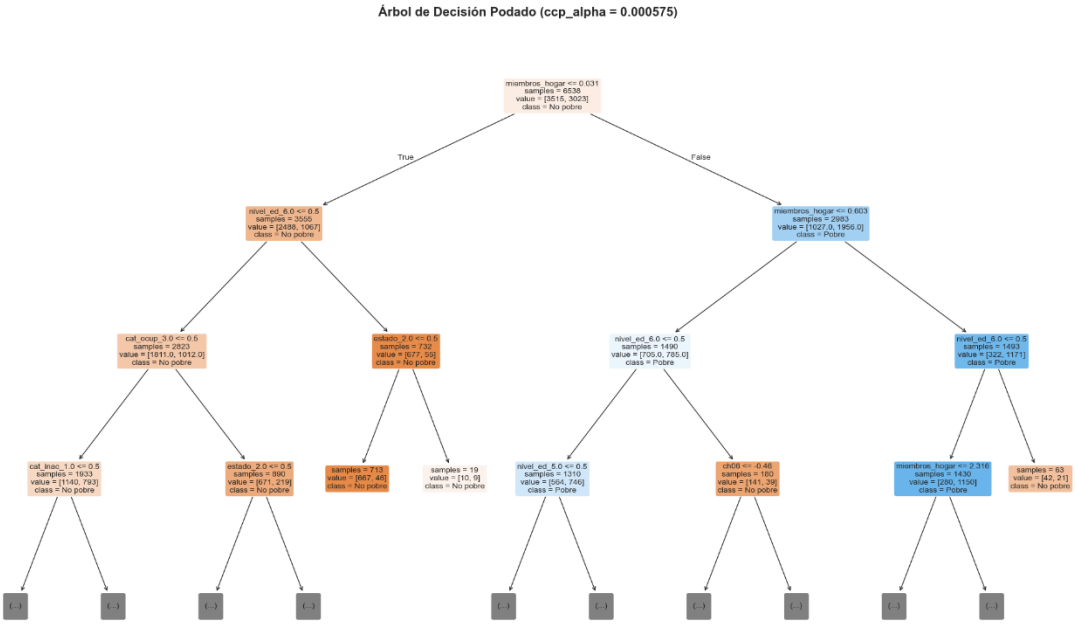
3.1 Estimación de un árbol de decisión podado (CART)

Para determinar el nivel óptimo de poda en un árbol de decisión, en esta sección utilizamos el hiperparámetro de costo de complejidad (ccp_alpha), que controla la penalización por complejidad del árbol. Un valor mayor de ccp_alpha produce árboles más simples (mayor poda); por el contrario, valores cercanos a cero permiten estructuras más profundas. Estimamos 546 árboles diferentes luego de filtrar valores únicos y no negativos de una grilla inicial de 791 valores, cada uno asociado a un valor de ccp_alpha único extraído del proceso de coste-complejidad del árbol sin podar. Para cada uno de estos árboles calculamos el error promedio de clasificación mediante validación cruzada de 10 folds, identificando un óptimo en ccp_alpha de 0.000575, con un error de clasificación mínimo de 0,2866, un rango de errores en la grilla de 0,2866 – 0,3499, y un accuracy de entrenamiento de 72.87%

El gráfico CV vs *ccp_alpha* muestra un patrón típico en árboles de decisión. Para ccp_alpha ≈ 0 (sin poda), el error de CV es relativamente alto (≈ 0,35), lo que evidencia sobreajuste (el árbol sin restricciones se adapta excesivamente a las particularidades del entrenamiento). A medida que aumenta ccp_alpha, el error cae muy rápidamente hasta alcanzar un mínimo alrededor de 0,000575, donde el árbol logra un mejor balance entre complejidad y capacidad predictiva. Por último, para valores mayores de ccp_alpha (>0.001) el error comienza a aumentar nuevamente, lo que indica que la poda excesiva simplifica demasiado el árbol, produciendo un ajuste subóptimo.

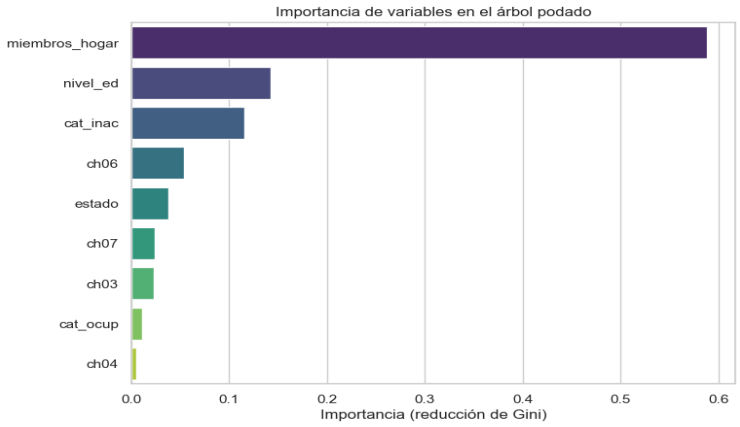


3.2Visualización del árbol podado por cross-validation



Con el ccp_alpha óptimo se estimó el árbol final. Las clases predichas son pobre y no pobre. Las características del árbol óptimo son las siguientes: a) Profundidad: 8, b) Número de nodos: 65, c) Número de hojas: 33 y d) Accuracy en entrenamiento: 0,7287

El árbol podado muestra una estructura clara y razonable. La primera división (la más informativa para separar las clases) es el número de miembros en el hogar. La segunda variable más utilizada en niveles superiores es máximo nivel educativo del hogar (nivel_ed_6.0). Luego aparecen divisiones asociadas al tipo de inactividad (cat_inac), a la condición de actividad (estado), a la edad y estado civil (ch06, ch07) y a la categoría de ocupación (cat_ocup). La lógica del árbol es consistente con los patrones de pobreza: el tamaño del hogar y nivel educativo generan divisiones gruesas entre pobres y no pobres, mientras que las categorías laborales refinan los nodos inferiores.El siguiente gráfico muestra la importancia relativa de las variables por reducción total de Gini:



Como puede observarse, miembros_hogar domina fuertemente el árbol (≈ 59% de la importancia total). Esto es coherente con la división raíz y refleja un predictor altamente discriminante para la condición de pobreza. El nivel educativo y la condición de inactividad también tienen pesos significativos, lo cual coincide con los determinantes socioeconómicos esperados. Variables como ch04 y cat_ocup tienen un peso muy bajo (<1%), lo que sugiere que aportan poca capacidad de discriminación en un modelo no lineal basado en particiones.

Obsérvese que, aunque LASSO no eliminó variables en la penalidad óptima, cuando la penalidad era más fuerte eliminaba primero las mismas variables que el árbol asigna menor importancia. Esto indica coherencia entre ambos métodos y refuerza la interpretación de que ciertas variables aportan mucha menos información que otras.

4. Comparación entre métodos

4.1 Desempeño comparativo de modelos

En esta sección evaluamos el desempeño predictivo de cinco modelos: 1) Regresión logística sin penalización, 2) Logit con LASSO (λ óptimo), 3) Logit con Ridge (λ óptimo), 4) Árbol de decisión podado (ccp_alpha óptimo) y 5) KNN con K óptimo por CV. Para todos ellos se calcularon las métricas principales: matriz de confusión, accuracy, recall de la clase "pobre", precisión, F1-score y AUC-ROC. La table 2 muestra que los modelos logísticos logran el major balance entre accuracy (71.27% – 71.31%) y AUC-ROC (0.7907-0.7908), con la menor tasa de error (1-accuracy = 28.69% - 28.73%). Los métodos no lineales (KNN y árbol) presentan performance inferior en todas las métricas.

Tabla 2. Desempeño general de los modelos

	Modelo	Accuracy	1- Accuracy	AUC ROC
1	Logit sin penalidad	71.31%	28.69%	0.7907
2	LASSO	71.27%	28.73%	0.7908
3	Ridge	71.27%	28.73%	0.7908
4	KNN	70.52%	29.48%	0.7790
5	Árbol podado	70.06%	29.94%	0.7767

La table 5 muestra que el Logit sin penalidad maximiza los verdaderos positivos (853) minimizando falsos negativos (443), lo cual resulta central para no excluir hogares pobres. Los métodos no lineales incrementan ambos tipos de error.

Tabla 3. Resultados de matrices de confusión

Modelo	Verdaderos positivos	Falsos positivos	Falsos negativos	Verdaderos negativos
Logit sin penalidad	853	361	443	1145
LASSO	851	360	445	1146
Ridge	851	360	445	1146
Árbol podado	826	369	470	1137
KNN	823	353	473	1153

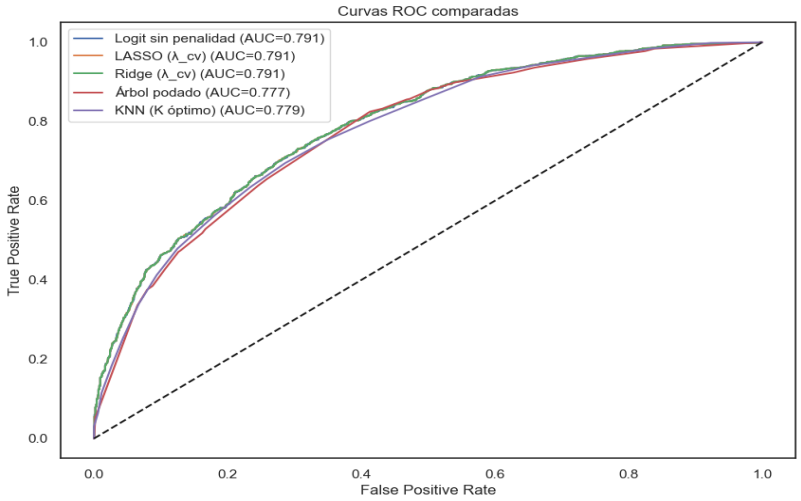
Comparando los errores en las matrices de confusion, se observa que los modelos logit (sin penalidad, LASSO, Ridge) son muy similares entre sí, esto significa que lost tres modelos logit equilibran razonablemente bien los dos tipos de error.Por su parte, el árbol podado aumenta tanto los falsos negativos (470) como los falsos positivos (369). Es decir, es un modelo menos preciso y menos sensible (con menor AUC y accuracy). Por último, el KNN empeora con respecto al logit la cuenta de falsos negativos (473) pero mejora la de verdaderos negativos (1153). Esto significa que KNN clasifica mejor a los no pobres, pero pierde sensibilidad en la clase pobre.

El mejor modelo para identificar pobres (maximizar verdaderos positivos) es el Logit sin penalidad (65.8%).

Tabla 4. Capacidad de identificación de pobres (Recall)

Modelo	Recall pobres
Logit sin penalidad	0,6582
LASSO	0,6566
Ridge	0,6566
Árbol podado	0,6373
KNN	0,6350

Todos los modelos tienen AUC cercana a 0.78–0.79, con diferencias muy pequeñas. El mejor AUC lo logran LASSO/Ridge, pero la diferencia con logit es irrelevante desde el punto de vista predictivo.



Como conclusion general, podemos decir que los métodos no lineales no mejoran la performance predictiva. Logit, LASSO y Ridge obtienen las mejores métricas en conjunto (con diferencias mínimas entre sí); mientras que el árbol podado y KNN, pese a captar relaciones no lineales, no superan al logit. Esto significa que la estructura del problema es mayormente lineal y que los modelos más flexibles no aportan ganancias significativas. Pero además, los modelos lineales (Logit, LASSO, Ridge) se caracterizan por una mayor interpretabilidad y comunicabilidad (coeficientes con signo y magnitud claros, fáciles de justificar ante decisores públicos). Por lo tanto, no quedan dudas de la superioridad de los modelos lineales en el contexto del presente problema.

4.2 Selección del mejor modelo

Dado que el Ministerio de Capital Humano busca identificar correctamente a personas pobres, priorizamos como métrica de evaluación el recall de la clase pobre. La idea es minimizar los falsos negativos, para evitar dejar sin ayuda a los grupos vulnerables. Por lo tanto, en vista de los resultados reportados, el mejor modelo para asignar recursos escasos a los más necesitados es la regresión logística sin penalización (lo que coincide con nuestra decisión previa).