Trabajo práctico Nº4

Ciencia de Datos

Grupo Nº4

Alumnos

Juan Augusto Alvarez Simonassi , Camila Gioja , María Delfina González Elosú

Profesores

María Noelia Romero

Ignacio Anchorena

Tomas Enrique Buscaglia

GitHub: Link

Parte 0: Datos utilizados y preparación de la base

En este trabajo se utiliza la base de datos de la Encuesta Permanente de Hogares (*EPH*) del año 2025, restringida a los hogares de la región Cuyo (región 42), dando continuidad al Trabajo Práctico 3. Se retoma la definición de la variable dependiente *POBRE*, construida a partir de la comparación entre el Ingreso Total Familiar (*ITF*) y la línea de pobreza ajustada por adulto equivalente, así como las principales variables explicativas previamente utilizadas, que incluyen características del hogar (*IX_TOT, adulto_equiv*), educación (*EDUC*), sexo (*MUJER*), condición laboral (*horastrab*, *JUBILADO*) y cobertura de salud (*PREPAGA*, *SIN_COBERTURA*). Las variables fueron transformadas a formato numérico y se realizó un control de valores faltantes e inconsistentes. A diferencia del TP3, el análisis del presente trabajo se concentra exclusivamente en el año 2025.

Parte A: Modelo de Regresión Logística con Regularización: Ridge y LASSO.

Con el objetivo de analizar cómo afectan las penalizaciones al comportamiento de los coeficientes del modelo logístico logrado anteriormente en el TP3, se estimaron regresiones logísticas penalizadas utilizando LASSO (L1) y Rigde (L2). Para ello se consideró una grilla de penalización definida como $\lambda=10^n$ con $n \in [-5,5]$. Dado que en *sklearn* la fuerza de penalización se controla mediante el parámetro $C = 1/\lambda$, se evaluaron once valores de C correspondientes a dicha grilla.

En la **Figura 1** se muestran las trayectorias de los coeficientes de la regresión logística bajo la regularización de LASSO y Ridge. En el **Panel A**, el caso de Lasso, se observa que a medida que aumenta λ los coeficientes se reducen en magnitud y muchas curvas convergen a cero. Este comportamiento refleja el efecto de *sparsity* característico de la penalización L1; el modelo tiende a "apagar" variables completas, fijando su coeficiente en cero cuando considera que no aportan capacidad predictiva adicional. Por el contrario, en el **Panel B** correspondiente a Ridge, los coeficientes también disminuyen con valores más altos de λ, pero prácticamente ninguno alcanza el valor cero. La penalización L2 distribuye el *shrinkage* entre todos los coeficientes de manera más suave, reduciendo sus magnitudes pero manteniendo todas las variables dentro del modelo.

Esta diferencia implica que LASSO tiende a seleccionar un subconjunto más reducido de predictores relevantes para explicar la pobreza, eliminando aquellos que no aportan información adicional, mientras que Ridge conserva todas las variables, suavizando sus efectos. Como resultado, Ridge tiende a producir modelos más estables frente a problemas de colinealidad entre predictores, mientras que LASSO favorece interpretabilidad y simplicidad al realizar selección automática de variables.

A continuación, para seleccionar el nivel adecuado de penalización de los modelos LASSO y Ridge, se utilizó *LogisticRegressionCV* con validación cruzada de 5 particiones. Para cada valor de la grilla $\lambda = 10^n$, con $n \in [-5,5]$, se estimó el error de clasificación medido como 1 - *accuracy* en el conjunto de validación. En la **Figura 2** se presentan los *boxplots* del error de clasificación para cada nivel de penalización, resumiendo la distribución del error en las cinco particiones de *cross-validation*.

En el **Panel A**, correspondiente a LASSO (L1), el error se mantiene relativamente bajo y estable para penalizaciones pequeñas ($\lambda \leq 10^0$). A partir de $\lambda \approx 10^2$ el error de validación tiene un aumento brusco del error y, para valores más grandes ($\lambda \geq 10^3$), se observa una degradación importante del desempeño, lo que refleja una sobre–penalización del modelo. El λ óptimo seleccionado por validación cruzada se ubica en la parte intermedia de la grilla, donde el error es mínimo. En el **Panel B**, correspondiente a Ridge (L2), el modelo presenta una trayectoria similar, aunque con un patrón más suave y progresiva. Para valores pequeños de λ el error se mantiene bajo, mientras que las penalizaciones más grandes producen un aumento progresivo del error. Nuevamente los valores extremos de λ deterioran la performance, mientras que el λ óptimo aparece en la región intermedia, indicando el equilibrio entre sesgo y varianza.

Se realizó, además, un gráfico de la proporción de coeficientes exactamente iguales a cero para cada valor de penalización en LASSO (**Figura 3**). Vemos que para valores pequeños, el modelo L1 mantiene casi todas las variables activas, aunque a medida que λ crece, esta proporción aumenta de forma monótona. Alrededor de $\lambda = 10^1$ inicia la eliminación de predictores y, para $\lambda \leq 10^5$, prácticamente todos los coeficientes son anulados. Esto confirma el efecto de selección automática de variables característico de la penalización L1, que induce *sparsity* en el modelo.

Posteriormente, en el inciso 3, se compararon los coeficientes estimados por la regresión logística sin penalización con aquellos obtenidos mediante regularización L1 (LASSO) y L2 (Ridge), utilizando en estos dos últimos los valores óptimos de λ seleccionados por validación cruzada. Dichos valores resultaron $\lambda = 1$ para el modelo de LASSO (L1) y $\lambda = 10$ para el modelo de Ridge (L2), ubicados en la región intermedia de la grilla $\lambda = 10^n$, lo que refleja un equilibrio entre sesgo y varianza. La **Tabla 1** presenta los coeficientes estimados para cada variable bajo los tres modelos.

En términos generales, los coeficientes de los modelos penalizados presentan una menor magnitud que los del modelo sin penalización, en línea con el efecto de *shrinkage* propio de las técnicas de regularización. Por ejemplo, la variable *PREPAGA* presenta un coeficiente de 0,905 en el modelo sin penalizar, que disminuye a 0,823 bajo LASSO y a 0,469 bajo Ridge; es decir, los coeficientes se “empujan” hacia cero, reduciendo la varianza del modelo y mitigando el sobreajuste. Asimismo, LASSO tiende a achicar con mayor intensidad los coeficientes de menor aporte predictivo, funcionando también como un mecanismo de selección de variables, mientras que Ridge conserva todas las variables en el modelo, aunque con coeficientes reducidos. Esto se observa, por ejemplo, en *MUJER*, cuyo coeficiente pasa de 0,129 en el modelo sin penalización a

0,026 en LASSO y a −0,208 en Ridge, lo que sugiere que el efecto asociado al género pierde robustez al controlar por el resto de los predictores. En línea con esto, Ridge reduce la magnitud de los coeficientes pero no los elimina: todos los predictores mantienen un coeficiente distinto de cero, a menudo muy pequeño. Si bien LASSO tiende a fijar algunos coeficientes cercanos a cero, la penalización L2 distribuye el *shrinkage* de forma más uniforme y no genera *sparsity*. En conjunto, la regularización permite aislar relaciones más estables y confiables, reduciendo la sensibilidad del modelo a la muestra y a la colinealidad entre variables.

Parte B: Árboles

Para la estimación del árbol de decisión se construyó inicialmente un árbol CART sin restricciones, a partir del cual se obtuvo la grilla de valores del hiperparámetro de complejidad *ccp_alpha* utilizando el método de *cost_complexity_pruning*. Luego, para cada valor de dicha grilla, se estimó un árbol podado y se evaluó su desempeño mediante validación cruzada de 10 particiones (*10–fold cross-validation*), calculando como métrica de desempeño el error de clasificación (1−accuracy).

La **Figura 4** muestra el error de validación (1 - *accuracy*) en función de *ccp_alpha*, representado en escala logarítmica. Se observa una forma aproximadamente en "V"; para valores pequeños de α el árbol presenta una complejidad elevada y un error relativamente alto, lo que refleja sobreajuste. En el extremo opuesto, para valores grandes de α el árbol se poda en exceso y pierde capacidad predictiva, lo que se traduce en un aumento del error por subajuste. El mínimo del error de clasificación se alcanza en torno a $\alpha = 0.001762$, valor que fue seleccionado como el óptimo por validación cruzada. En consecuencia, el árbol utilizado en los siguientes incisos corresponderá al árbol podado con dicho alpha, ya que representa el mejor balance entre complejidad del modelo y desempeño predictivo fuera de la muestra (sesgo-varianza).

Posteriormente se realizó la **Figura 5**, que muestra en el **Panel A** el árbol de decisión podado utilizando el valor óptimo de *ccp_alpha* = 0.001762, y en el **Panel B** la importancia de los predictores calculada a partir del árbol podado. Estos pesos se miden como la reducción acumulada de la impureza (índice de Gini) generada por cada variable a lo largo de todas las particiones del árbol. Vemos en el nodo raíz del árbol que el primer criterio de partición que corresponde a la variable *SIN_COBERTURA,* lo que indica que la falta de cobertura de salud es el factor más relevante en la primera segmentación entre pobres y no pobres. Entre los individuos sin cobertura de salud, el árbol profundiza la clasificación a partir de la variable *IX_TOT*, lo que sugiere que una vez considerada la cobertura, la cantidad de personas viviendo en el mismo hogar es un determinante central en la identificación de la pobreza. En el caso de quienes sí cuentan con cobertura, el siguiente criterio de partición es *horastrab*, lo que sugiere que más horas de trabajo contribuyen a diferenciar entre situaciones de pobreza y no pobreza en el grupo. Por último, para ramas más específicas aparecen umbrales asociados a *EDUC*, y *PREPAGA*, que refinan la clasificación.

En cuanto a los pesos de las variables, la variable con mayor peso es *SIN_COBERTURA*, seguida por *IX_TOT* y *horastrab,* mientras que *JUBILADO, MUJER* y *adulto_equiv* muestran importancias prácticamente nulas. Este patrón es, en general, consistente con los resultados obtenidos mediante regularización, ya que fueron las variables más penalizadas por LASSO en la regresión logística. En este sentido, ambos enfoques (árboles de decisión y regularización L1) señalan de forma consistente cuáles variables aportan escasa información adicional para la clasificación de pobreza.

Parte C: Comparación entre métodos

En la comparación entre los distintos métodos se evaluó el desempeño predictivo de los modelos estimados sobre el conjunto de test 2025: regresión logística sin penalización, regresión logística con penalización L1 (LASSO), regresión logística con penalización L2 (Ridge), árbol de decisión CART podado y un clasificador k-NN con K seleccionado mediante validación cruzada. La evaluación se realizó a partir de la matriz de confusión, la *accuracy*, la sensibilidad de la clase "pobre" y las curvas ROC junto con el área bajo la curva (AUC), presentadas en la **Figura 6**, utilizando un umbral de clasificación de 0.5.

Los tres modelos logísticos presentan un desempeño prácticamente idéntico. El logit sin penalización alcanza una *accuracy* de 0.730 y un AUC de 0.776, mientras que el modelo LASSO obtiene una *accuracy* de 0.731 y un AUC de 0.777. De manera análoga, el modelo Ridge presenta una *accuracy* de 0.730 y un AUC de 0.777. Estos resultados indican que la incorporación de penalización no genera mejoras sustantivas en términos de desempeño predictivo fuera de muestra, aunque sí aporta ventajas en términos de estabilidad de los coeficientes, tal como se analizó en los incisos previos. El árbol de decisión CART podado alcanza la *accuracy* más alta entre los modelos considerados (0.741), aunque su AUC es ligeramente inferior (0.765). Por su parte, el clasificador k-NN con K seleccionado por validación cruzada presenta una *accuracy* de 0.735 y un AUC de 0.769, ubicándose en un nivel de desempeño intermedio entre los modelos logísticos y el CART.

En conjunto, todos los modelos presentan valores de AUC claramente superiores a 0.5, lo que evidencia una capacidad de discriminación moderada entre hogares pobres y no pobres. No se observa, sin embargo, un dominador claro en términos de desempeño predictivo global: los modelos logísticos concentran los mayores valores de AUC, mientras que el CART maximiza levemente la *accuracy*. Las diferencias entre modelos son acotadas, lo que sugiere que el desempeño predictivo es relativamente estable entre las distintas técnicas consideradas.

Por último, para el último inciso, dado que el Ministerio de Capital Humano está interesado en identificar a los grupos vulnerables para dirigir los recursos de su programa de alimentos, las métricas más relevantes no van a ser la accuracy global ni el AUC, sino principalmente la sensibilidad (*recall*) de la clase "pobre" y la cantidad de falsos negativos (FN). Esto se debe a que los falsos negativos presentan hogares pobres que serían erróneamente clasificados como no pobres y por lo tanto, quedarían excluidos del programa, constituyendo un error socialmente costoso. A partir de la **Tabla 2** podemos ver que el árbol de decisión CART presenta la mayor sensibilidad (0.518) y la menor cantidad de falsos negativos (214) en comparación con los modelos logísticos y el KNN. En contraste, si bien los modelos logísticos muestran valores más altos de AUC (0.777), su sensibilidad es inferior, además que la diferencia de AUC no es tan grande. En consecuencia, la respuesta sobre cuál es el "mejor" modelo cambia si consideramos el objetivo de política pública: mientras que desde un criterio puramente estadístico los modelos logísticos presentan una mejor capacidad discriminante global, desde la perspectiva de asignación de recursos a los

más necesitados el modelo CART resulta el más adecuado, ya que minimiza la exclusión de hogares pobres del programa.

Anexo

Figura 1

Trayectoria de coeficientes

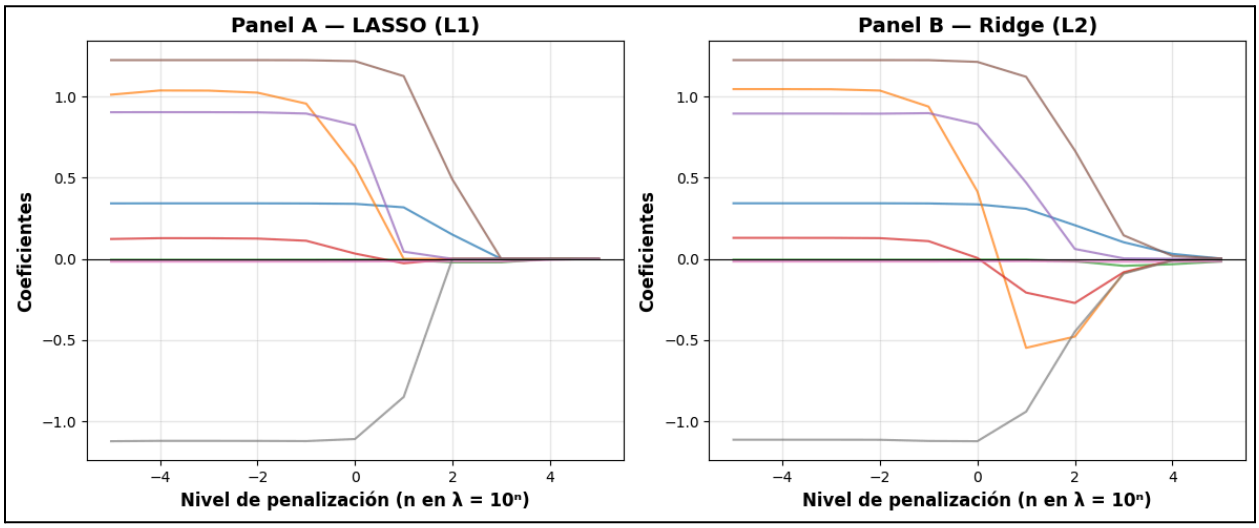
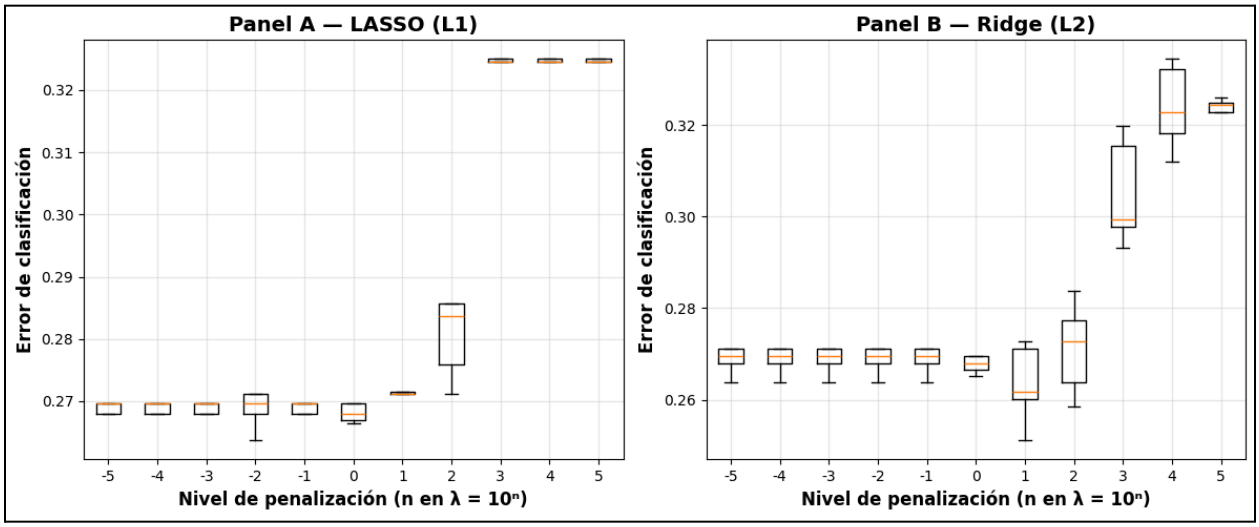


Figura 2

Distribución del error por λ



Nota: el error de clasificación fue medido como 1 - *accuracy*

Figura 3

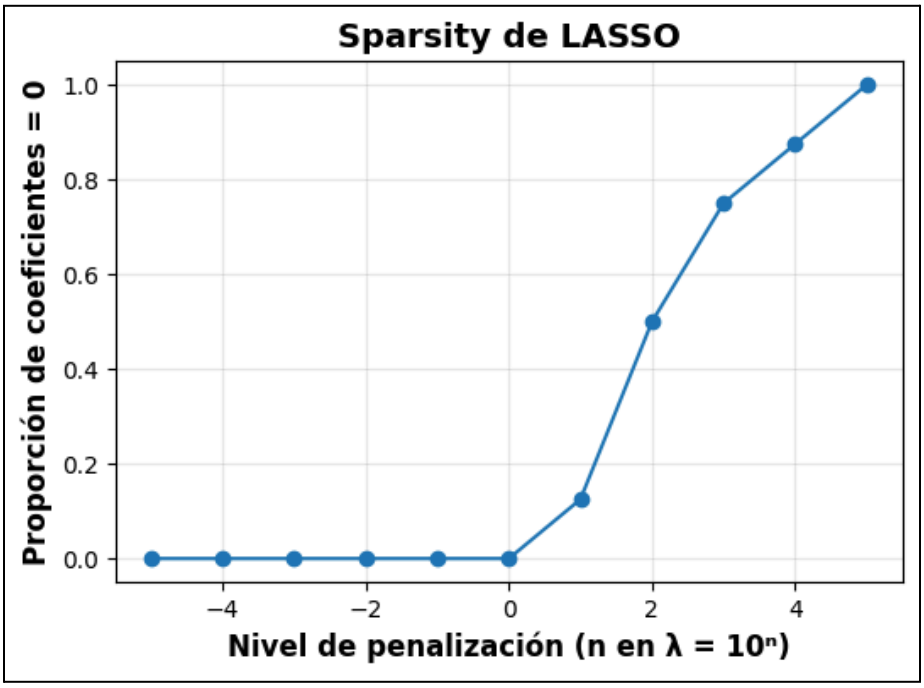


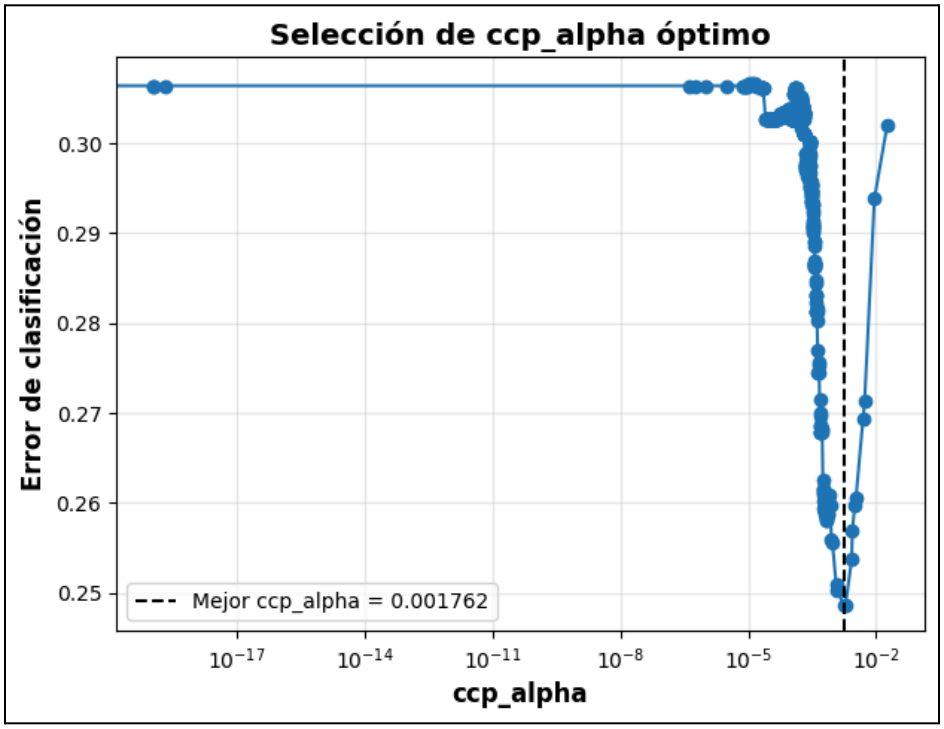
Tabla 1

Coeficientes estimados para cada variable por los modelos de regresión.

	Modelo de regresión		
Variable	**Logit sin penalidad**	**LASSO (L1)**	**Ridge (L2)**
Intercepto	-3.242	-2.761	-1.555
IX_TOT	0.342	0.339	0.308
adulto_equiv	1.049	0.547	-0.548
EDUC	-0.006	-0.005	-0.005
MUJER	0.130	0.026	-0.208
PREPAGA	0.905	0.823	0.469
SIN_COBERTURA	1.223	1.218	1.121
horastrab	-0.017	-0.016	-0.015
JUBILADO	-1.120	-1.111	-0.941

Nota: Las columnas LASSO (L1) y Ridge (L2) utilizan valores óptimos de penalización seleccionados por validación cruzada (λL1 = 1, λL2 = 10).

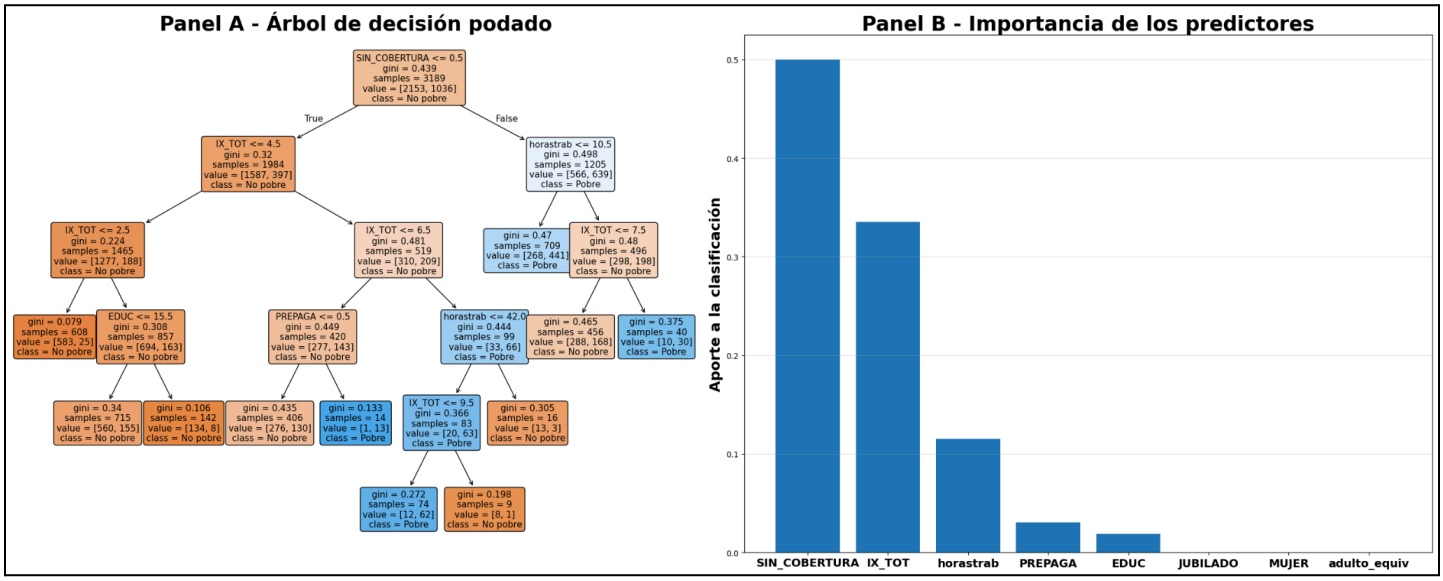
Figura 4



Nota: El error de clasificación fue medido como 1-accuracy y el eje x que mide *ccp_alpha* se presenta en escala logarítmica. La línea punteada indica el valor de *ccp_alpha* que minimiza el error medio de clasificación obtenido mediante validación cruzada de 10 particiones.

Figura 5

Árbol de decisión junto con aporte relativo de cada variable



Nota. La importancia de cada predictor se calcula como la reducción acumulada de la impureza (índice de Gini) generada por los splits del árbol de decisión.

Figura 6

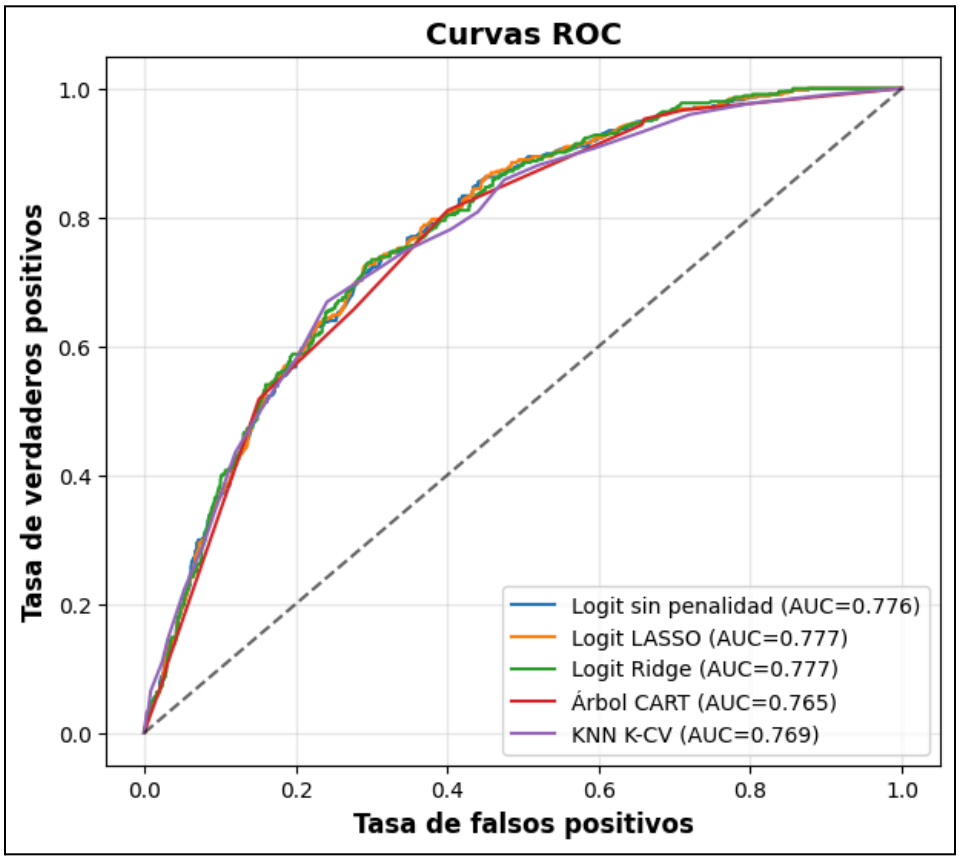


Tabla 2

Desempeño predictivo de los modelos para la identificación de hogares pobres

Modelo	Accuracy	Sensibilidad	FN	AUC
Logit sin penalidad	0.730	0.437	250	0.776
Logit LASSO (L1)	0.731	0.414	260	0.777
Logit Ridge (L2)	0.730	0.412	261	0.777
Árbol CART podado	0.741	0.518	214	0.765
KNN con K-CV	0.735	0.486	228	0.769

Nota: La sensibilidad corresponde al recall de la clase "pobre". FN = falsos negativos (hogares pobres clasificados como no pobres=.